2013

Molecular phylogenetics of the North American snake tribe Thamnophiini

John David McVay
Louisiana State University and Agricultural and Mechanical College

MOLECULAR PHYLOGENETICS OF THE NORTH AMERICAN SNAKE TRIBE THAMNOPHIINI

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Biological Sciences

by
John David McVay
B.S., University of Texas, 2001
M.S. Texas Tech University, 2007
August 2013

ACKNOWLEDGMENTS

First and foremost, my dissertation would not have been possible without the guidance and support of my advisor, Dr. Bryan Carstens. My understanding of evolution and of science has advanced immensely under his tutelage, and he instilled in me the confidence and tools to approach methodological challenges. I will always be indebted to the opportunity I was given.

My committee members, Drs. Christopher Austin, Robb Brumfield and James Cronin, each provided encouragement, guidance and support through their respective experiences and expertise. I am particularly indebted to Dr. Austin for hiring me as a technician in his laboratory; collaboration with him and discussion with faculty, staff, and students at the time ultimately encouraged me to return to graduate school. I am also grateful for the opportunity to collaborate with Dr. Oscar Flores-Villela, who provided access to samples from Mexico, encouragement, support, and expertise in Mexican herpetology.

My understanding of the field would not be where it is without frequent, amazing discussions with members of the Carstens lab. Sarah Hird, Noah Reid, Jordan Satler, Tara Pelletier, Dr. Maggie Koopman, Dr. Erica Tsai, Daniel Ence, Danielle Fuselier and Dr. Amanda Zellmer each provided their input from their unique perspectives of expertise to make conversations and research more fruitful. Additionally, the many pleasant and helpful discussions with Dr. Jeremy Brown doubtlessly improved my dissertation and approach to research.

Much of my time as a graduate student was devoted to laboratory-based data collection. This work was facilitated with aptitude and amenability by Dr. James Maley, Dr. Scott Herke, Dr. Nathan Jackson, Donna Dittman and Eric Rittmeyer.

I would not have been able to make it through the rigors of graduate school with out support from my friends. Dr. Ron Eytan, Dr. James Maley, Jordan Satler, and Tara Pelletier will stand out in my memory among many. Noah Reid and Sarah Hird were and are as close friends to me as could be. Noah is an always-reliable thinking, talking, laughing, birding, and biking companion. Finally, Ricky Rodriguez provided unconditional friendship, support, access to pets, wit, tongue-in-cheek criticism, and a place to hang my hat during my matriculation.

My parents, Drs. Catherine and Ted McVay provided constant, unconditional love and support during this endeavor and throughout my life. Their hard work and pursuit of knowledge also provided the inspiration to pursue a career in academia.

Finally, I could not have completed my dissertation without the love and encouragement of my wife, Megan McVay. Her success inspires me, her hard work amazes me, her love supports me.

TABLE OF CONTENTS

ABSTRACT

Advancements in phylogenetic theory and methodology coupled with improvements in computational and sequencing technology facilitate study of the divergence and diversification patterns of life. I apply our current understanding to further explore the relationships and evolution of the North American snake tribe Thamnophiini, as well as to address current topics in phylogenetic and taxonomic methodology.

There are two paradigms for the phylogenetic analysis of multi-locus sequence data: one which forces all genes to share the same underlying history, and another that allows genes to follow idiosyncratic patterns of descent from ancestral alleles. The first of these approaches (concatenation) is a simplified model of the actual process of genome evolution while the second (species-tree methods) may be overly complex for histories characterized by long divergence times between cladogenesis. Rather than making an a priori determination concerning which of these phylogenetic models to apply to our data, I seek to provide a framework for choosing between concatenation and species-tree methods that treat genes as independently evolving lineages. In Chapter 2 I demonstrate that parametric bootstrapping can be used to assess the extent to which genealogical incongruence across loci can be attributed to phylogenetic estimation error, and demonstrate the application of our approach using an empirical dataset from 10 species of the Natricine snake sub-family. Since our data exhibit incongruence across loci that is clearly caused by a mixture of coalescent stochasticity and phyogenetic estimation error, we also develop an approach for choosing among species tree estimation methods that take gene trees as input and those that simultaneously estimate gene trees and species trees.

Ideally, existing taxonomy would be consistent with phylogenetic estimates derived from rigorously analyzed data using appropriate methods. In Chapter 3 I present a multi-locus molecular analysis of the relationships among nine genera in the North American snake tribe Thamnophiini in order to test the monophyly of the crayfish snakes (genus *Regina*) and the earth snakes (genus *Virginia*). Sequence data from seven genes were analyzed to assess relationships among representatives of the nine genera by performing multi-locus phylogeny and species tree estimations, and we performed constraint-based tests of monophyly of classic taxonomic designations on a gene-by-gene basis. Estimates of species trees demonstrate that both genera are paraphyletic, and this inference is supported by a concatenated tree. This finding was supported using gene tree constraint tests and Bayes factors, where we rejected the monophyly of both the crayfish snakes (genus *Regina*) and the earth snakes (genus *Virginia*).

Progress in our understanding of molecular evolution necessitates a more thorough assessment of the phylogeny of thamnophiine snakes, whose relationships have not been fully resolved, and whose previous phylogenetic estimates are based solely on mitochondrial sequence data. In Chapter 4, I present the most data and taxa robust phylogenetic estimate of Thamnophiini to date, including 50 taxa and sequence data from 8 independently sorting loci. Our findings support the taxonomic recommendations proposed in Chapter 3. Additionally, I estimated the timing of divergence among the three

major lineages to have occurred during the Miocene period (~14-11MYA), with higher than expected diversification in the garter snaked during the Pliocene period (~2-6MYA). Finally, we demonstrate that prey choice is labile, and thus an unreliable character for phylogeny reconstruction.

Combined, these chapters present a thorough examination of the molecular phylogenetics of thamnophiine snakes. The novel methodological approaches may serve as a guideline for future research. Through estimating a robust phylogeny and suggesting taxonomic changes where appropriate, this work provides a foundation for phylogenetically-based studies of this group.

CHAPTER 1.
MOLECULAR PHYLOGENETICS, TAXONOMY AND THEIR APPLICATIONS TO THAMNOPHIINI

1.1. The path to modern taxonomy and phylogenetics

As part of what became the modern evolutionary synthesis (Huxley, 1942), taxonomists improved upon the Linnean classification system by striving to categorize organisms based on their evolution. They accomplished this through careful examination of their characteristics, culminating in taxonomy with an underlying phylogeny, often based on gestalt of organisms. In contrast to this classification paradigm, numerical taxonomy (Sneath and Sokal, 1973), sought to classify organisms based on overall similarity using a mathematical approach. Aided by nascent advancements in computing techonology, numerical taxonomy, or phenetics, promoted itself as an objective process compared to evolutionary taxonomy (Sneath and Sokal, 1973). The fundamental philosophical differences to these approaches culminated in raucous clashes between supporters of each methodology. While the debate among evolutionary and numerical taxonomists continued, Hennig (1950), rather than relying on overall similarity for classification, proposed that taxa be ordered into monophyletic groups, and relationships be based on shared homologous characters. This gave rise to the practice of cladistics and philosophy of parsimony: reconstructing phylogenies based on the fewest number of character state changes (i.e., minimizing homoplasy) across the length of the tree. This method incorporated assumptions about evolution, and thus was met with criticism from pheneticists, who did not believe that it was plausible for researchers to make inferences concerning process given the current understanding of evolution (for a review, see Hull [1990]). Eventually, cladistics was largely accepted as the preferred method of phylogeny estimation, and though distance-based methods are considered useful by some (e.g., Li 1997) for quickly estimated approximations of phylogeny, phenetecists' most important contribution was the implementation of mathematical and computer-based analyses of phylogenetic data (Sneath, 1995).

Where parsimony invokes analysis of shared (synapomorphic/symplesiomorphic) characters as criteria for classification, it does not assume to know anything about the underlying nature of character evolution. The likelihood function, pioneered by (Edwards and Cavalli-Sforza, 1964), and improved upon and facilitated by Felsenstein (1973, 1981) incorporates models of character state evolution into the phylogenetic inference in a way that estimates the parameters of character evolution across the depth of the tree. This distinction between the two cladistic methods has led to disagreements among scholars; for discussions of the debate see Felsenstien (2004). Later, likelihood methods were developed (Yang and Rannala, 1997) and implemented (Huelsenbeck and Ronquist, 2001) in a Bayesian framework, which allow researchers to incorporate prior information and the likelihood of the data to produce a posterior distribution of parameter estimates, including the phylogeny.

Concurrently, as tools to gather protein and later DNA sequence data emerged, models were developed to estimate the manner in which the molecules evolved. Jukes and Cantor

(1969) proposed a single rate of substitution among nucleotides; later, more complex models (e.g., Hasegawa et al., 1985; Kimura, 1980; Tamura and Nei, 1993) incorporated more rates and additional parameters to explain the mode of nucleotide evolution. Other important contributions include the tools created to choose between competing models of nucleotide substitution (e.g., Minin et al., 2003; Posada, 2008; Posada and Crandall, 1998), development of algorithms for multiple sequence alignment (for a review see Notredame, 2007), and development and incorporation of the relaxed molecular clock (allowing for a flexible rate of nucleotide evolution across branches of a phylogeny) into likelihood and Bayesian methods (Drummond et al., 2006). Many studies have been committed to testing and improving to each of these methodologies; their findings have culminated in the current state of molecular phylogenetics, a heavily computational and largely model-based field.

In addition to these computational and methodological advances in the field of molecular systematics in the last 20 years, there has been a paradigm shift in the approach to estimating species. Following Kingman's (1982) description of the coalescent model of population genetics, Maddison (1997) suggested that gene genealogies will not necessarily share the same topology as the underlying species tree. This fostered research culminating in the idea that concatenation of multiple, independently sorting loci may bias estimates of phylogeny, and in extreme cases may positively mislead researchers in their conclusions of relationships among lineages (Degnan and Rosenberg, 2006, 2009). I attempt to address these concerns in each of my research chapters.

Once produced, phylogenetic estimates can be applied to a variety of studies to test evolutionary hypotheses, such that comparative studies of ecology, development or behavior can be conducted in an evolutionary context. Through mapping discrete or continuous characters onto topologies, researchers can estimate the state of said characters across the tree in a statistically informative way (Felsenstein, 1985). This method has been implemented in a wide array of studies, including many focused on squamate reptiles. For example Brandley et al. (2008) assessed rates of limb loss among squamates; Pyron and Burbrink (2009) elucidated patterns of mimicry by non-venomous milksnakes (tribe Lampropeltini) of venomous coral snakes (family Elapidae). Sites et al. (2011) provides a review of the accumulated work on character evolution in squamates. By incorporating fossils as calibration points across a phylogeny, researchers can estimate times of divergence among taxa (e.g., BEAST [Drummond, et al., 2012]). Rates of diversification may fluctuate over time within a phylogeny, perhaps corresponding to geologic events or key innovations within a lineage. Changes in this rate can be detected, and hypotheses concerning diversifications can be tested by implementing tools designed to extrapolate rates from phylogenetic estimates (e.g., Paradis, 2004; Rabosky, 2008). To paint a broader picture of evolution within a clade, all of these methods can be combined to assess correlation between character evolution, timing of divergence and rates of diversification (for a review, see O'Meara, 2012).

1.2. The Thamnophiini

My group of study is part of a global distribution of snakes, Natricinae (Bonaparte, 1838). This subfamiliy of the Colubridae contains 28 currently-recognized genera, with representatives on all continents except Antarctica (Uetz, 2008). The majority of this radiation includes species that are associated with aquatic habitats, utilizing both still and moving water as foraging sites and refuges from predators. I focus my dissertation on the North American radiation of natricine snakes: the tribe Thamnophiini. This monophyletic assemblage (most recently Lawson et al., 2005) is represented by 9 extant genera, spanning from Canada to Costa Rica. Convergence of bauplans (gartersnake-like, watersnake-like, and even crayfish specialists) is evident within Thamnophiini, as well as across the global distribution. Evolutionary convergence has generated debate among molecular and morphological researchers concerning the taxonomy of this group, with differing conclusions as a function of whether researchers analyzed genetic or morphological data. One prominent example relates to the status of the genus Virginia (Lawson, 1985; Varkey, 1979), with morphological data supporting a monophyletic Virginia and molecular data suggesting paraphyly. Similar patterns are evident at broader scales, for example the most recent molecular estimate of phylogeny that includes representatives of the majority of lineages (Alfaro and Arnold, 2001) suggests three major lineages within Thamnophiini: the garter snakes (including the genera *Thamnophis* and *Adelophis*), the water snakes (including *Nerodia, Regina grahamii* and *septemvittata*, and *Tropidoclonion*) and the semi-fossorial clade, made up of *Clonophis*, *Seminatrix*, *Storeria*, *Regina alleni* and *rigida*, and *Virginia*. Each lineage exhibits different distribution patterns, with the water snakes and semi-fossorial snakes having peak diversity in the southeastern United States, and the garter snakes with multiple regions of concentrated diversity, in the Eastern and Western U.S., and México (figure 1.1). The primary aim of my dissertation is to resolve relationships within the Thamnophiini by collecting molecular data from multiple loci in order to infer relationships among these problematic taxa.

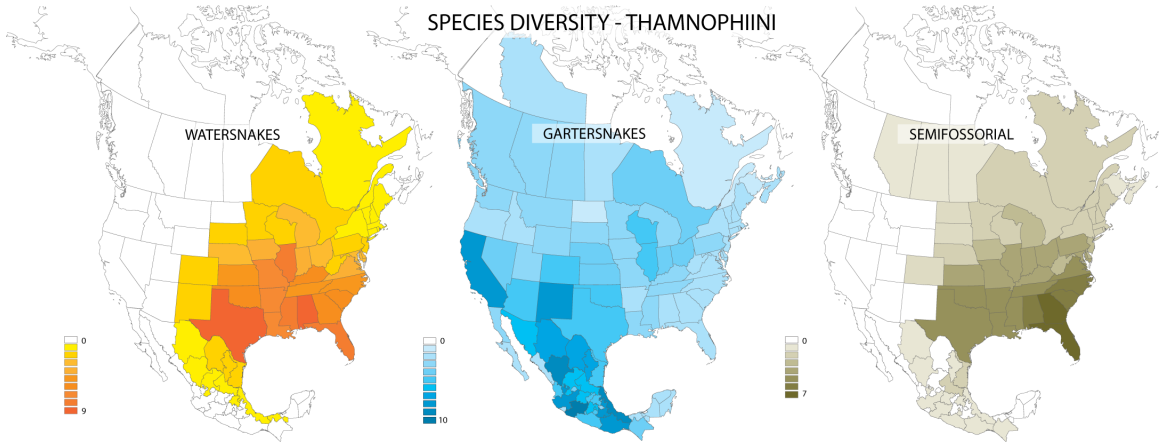


Figure 1.1. Number of species by state or province in Canada, Mexico and the United States, among the three major lineages of Thamnophiini. Lineages are denoted in the text.

1.3. Overview of Chapters

The focus of my dissertation combines two key components of my research interests: I am driven to approach the methodological aspects of my research from a well-informed and critical angle. I strive in my first two chapters to present a robust treatment of methods, serving to inform the methodological choices made for future research. Secondly, I have focused the bulk of my academic research on the thamnophiine snakes, in which I see a wealth of opportunities to explore evolutionary hypotheses. Additionally, despite their largely dull coloration and their off-putting nature when handled, I am drawn to these species with admiration and fascination. Each research chapter is written in journal-style, with references for all chapters compiled at the end of the document.

Chapter 2.—Given the advent of a newer model of species evolution (multi-species coalescent) and mounting evidence that concatenation may be positively leading in some cases, how do we best choose a model of evolution for a given dataset? I take an a priori approach to evolutionary model selection, with a discussion of performance among differing methods of species tree estimation (including Minimize Deep Coalescence [MDC, implemented in Mesquite; Maddison and Maddison, 2011], STEM [Kubatko et al., 2009], *BEAST [Drummond et al., 2012]), utilizing both simulated and empirical (including the thamnophiine genus *Nerodia*) sequence data. I also assess whether discordance among independently sorting loci using can be caused solely by phylogenetic estimation error.

Chapter 3.—When utilizing molecular sequence data for studies that span deep time, it can be challenging to collect data that inform us on the shared history of multiple lineages. Specifically, polymerase chain reaction (PCR) relies on two conserved genomic oligonucleotide ("primer") regions that flank target loci; as these loci evolve, the accumulation of substitutions carry information about relationships among lineages. Difficulty arises when substitutions occur in the primer regions, rendering this information inaccessible via PCR. Thus, many loci that can be amplified across deep time exhibit a low rate of evolution, and, consequently, the estimates derived from these loci are often poorly resolved (in the conluding chapter, I will discuss some recent technological advances that may serve to mitigate this challenge). How can I best apply molecular data when phylogenetic analyses yield less-than-ideal results? I address this question using Bayesian hypothesis testing, incorporating a recent method for estimating marginal likelihoods (stepping stone sampling; (Xie et al., 2011), which allows for a robust comparison between phylogenetic hypotheses, despite the lack of a well-resolved phylogenetic estimate. I apply these methods to lingering taxonomic disagreements and uncertainties within thamnophiine snakes, and make suggestions to amend the current taxonomy.

Chapter 4.—Incorporating novel genomic loci developed for this study with previously developed molecular markers, I estimate the most robust phylogeny of Thamnophiini to date, both in terms of number of loci and in breadth of taxonomic sampling. In order to gain insight into the nature of divergence and diversification within this group, I employ to the phylogeny to estimate diet and habitat preference across the history of their

radiation. I also assess the timing of divergence across the tree, and use statistical methods to characterize the rate of diversification through time. I discuss the taxonomic implications of chapters three and four.

Chapter 5.—Finally, I conclude with some comments synthesizing the ideas within each chapter, and discuss the current state of molecular phylogenetics and opportunities for further evolutionary research within the thamnophiine snakes. This work represents an attempt to contribute to the legacy of the molecular and morphological systematic work conducted at Louisiana State University, including the large body of research by Robin Lawson, Douglas Rossman, the late Herb Dessauer (1921-2013), and their many students and collaborators.

CHAPTER 2.
PHYLOGENETIC MODEL CHOICE: SPECIES TREES OR CONCATENATION?

2.1. INTRODUCTION

There are two primary paradigms for estimating phylogeny from multi-locus sequence data (Edwards, 2009). The conventional method, which developed from arguments in favor of total evidence (Kluge, 1989), estimates phylogeny by concatenating data across multiple genes collected from exemplar samples. In this approach, the data are treated as a single locus, and essentially the estimate of genealogy from each locus is averaged across genes. Underlying this method is the intuition that phylogenetic accuracy improves with an increase in the number of variable sites (Hillis et al., 1994). While this assumption certainly holds within a particular locus, applying this method across multiple loci requires the assumption that the gene trees across loci share a similar patterns of diversification. When this is demonstrably not the case, incongruence across loci is attributable to phylogenetic estimation error rather than to coalescent processes (e.g., the independent sorting of alleles across loci). Recently, the primacy of concatenation has been challenged on several fronts (Degnan and Rosenberg, 2006; Carstens and Knowles, 2007a; Kubatko and Degnan, 2007; Degnan and Rosenberg, 2009; Liu et al., 2009), and methods that estimate phylogeny while allowing for incongruence across loci due to coalescent processes have been proposed. These coalescent-based approaches to phylogeny inference estimate species tree either given gene trees (Kubatko et al., 2009; Maddison and Maddison, 2009), or estimate gene trees and species tree topologies simultaneously (Liu, 2008; Heled and Drummond, 2009). Either approach accounts for population-level processes, such as the incomplete sorting of ancestral polymorphism that can cause gene tree discordance.

Given the growing criticism of concatenation, empiricists are faced with a vexing decision regarding the choice of phylogenetic method to apply to their system. Coalescent-based approaches are often favored a priori in phylogeographic investigations, where the incomplete sorting of ancestral polymorphism can be dramatically evident across loci (Carstens and Knowles, 2007a; Knowles and Carstens, 2007; Brumfield et al., 2008; Godinho et al., 2008; King and Roalson, 2009; Leache, 2009), while concatenation continues to be favored among those working at deeper taxonomic levels (Li and Orti, 2007; Wiens et al., 2008; Blanga-Kanfi et al., 2009;). However, it is clear that population level processes such as the sorting of ancestral polymorphism have occurred throughout the history of life; further, one of the central theses of the modern synthesis is the expectation that evolutionary processes within populations ultimately produce phylogenetic patterns (Simpson, 1944). This led Edwards (2009) to argue that species tree approaches are preferable on first principles. Philosophical implications aside, the question of phylogenetic method choice is also of dramatic practical importance because the ideal sampling schemes for concatenation and coalescent-based approaches are quite different. Since the former assumes that population-level processes do not have an effect on phylogeny estimation, systematists

who concatenate their data benefit from sampling as many genes as possible and fewer individuals per species. Alternatively, coalescent-based approaches appear to be most accurate with intermediate numbers of loci and multiple individuals sampled within species (Maddison and Knowles, 2006; McCormack et al., 2009). This places an empiricist in a difficult position; optimally they need to recognize which of these approaches appears to be appropriate given their data before all of it is collected in order to employ the optimal sampling scheme. In this study we propose an approach to answering this question using a preliminary data set of 7 genes from 2 individuals for each of 10 species of thamnophiine snakes. Given our data, how should we determine which of the competing phylogenetic paradigms to utilize?

Perhaps the most important evidence available to empiricists who seek to objectively determine whether to concatenate their data or use species-tree methods is the degree of incongruence among loci. If the gene trees are mostly congruent, this is evidence that the branch lengths of the species tree are sufficiently long to have allowed lineage sorting to reach completion, and thus concatenation may be justified. Alternatively, incongruence among gene trees may be caused by coalescent processes and would suggest that coalescent-based methods are required. One approach would simply be to measure the incongruence across gene trees using a metric for tree comparison such as the Robinson-Foulds distance (Robinson and Foulds, 1981). Distributions of the pairwise RF distances can be substantial at shallow phylogenetic depths; this incongruence can also persist at deeper levels (Figure 2.1). However, observed discordance among gene tree estimates can arise from other neutral sources such as mutational stochasticity, as well as phylogenetic estimation error, and thus a major challenge for empiricists is determining if the observed incongruence across gene trees can be attributed to phylogenetic estimation error alone. It is reasonable to conclude that concatenation is appropriate when the level of discord is of a magnitude that can be attributed to phylogenetic estimation error, here the substitutions across loci will provide valuable information regarding ancestral nodes. Conversely, gene tree estimates that are incongruent to a greater extent than would be expected due to phylogenetic estimation error alone is an indication that coalescent uncertainty has caused the discord, and therefore must be accounted for through the use of species tree estimation approaches. Here we use parametric bootstrapping to conduct a series of pairwise tests to ascertain whether the incongruence across genealogies estimated from our empirical data can be attributed to phylogenetic estimation error alone.

We explore these questions using data collected from the North American colubrid snake tribe Thamnophiini, which consists of ~58 species representing nine genera. Previous studies (Alfaro and Arnold, 2001; de Queiroz et al., 2002) have estimated partial phylogenies of this group with differing results, yet to date no complete phylogeny has been estimated. Most recently, Alfaro (2003) published a Bayesian estimate of phylogeny in which the *Nerodia*, the water snakes, were not monophyletic. Specifically, two species of the genus *Regina* and the monotypic *Tropidoclonion* were nested within *Nerodia*. While this result was not strongly supported by the data, the relationships within the Thamnophiini, remain unsatisfactorily resolved. While our ultimate research goal is to

resolve the phylogeny of this group, our proximate goal is to determine which phylogenetic method is appropriate so that we can identify the optimal sampling scheme.

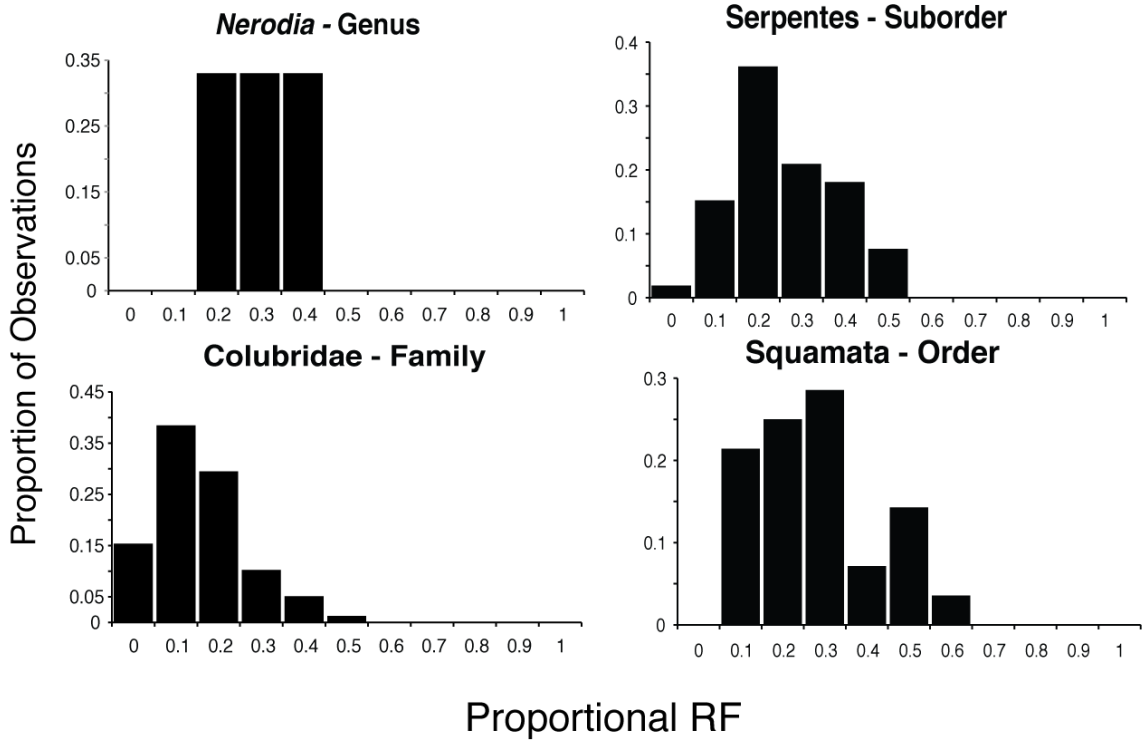


Figure 2.1. Histograms of pairwise symmetric difference distance among gene trees from four multi-locus empirical datasets. To assess the degree to which individual gene trees share the same topology for multigene datasets representing four depths of phylogeny (Table 2.1), we used the symmetric difference distance (RF distances; Robinson and Foulds, 1981) to compare all trees in a pairwise fashion using PAUP*. Observations of zero indicate no topological incongruence between two trees.

2.2 METHODS

2.2.1 Empirical

Data collection.—Molecular sequence data from five nuclear and two mitochondrial gene fragments were collected from seven species of the snake genus *Nerodia*, two North American relatives (*Regina grahamii* and *Tropodoclonion lineatum*), and an old world relative (*Natrix natrix*). The additional North American species were included to include an individual that has been previously suggested to be within *Nerodia* (*R. grahamii*) and to include one from a putatively deeper split (*Tropidoclonion*; Alfaro, 2003). DNA was extracted from tissue or blood using DNEASY kit (QIAGEN, Hilden, Germany) following manufacturer's protocols. Each fragment was amplified via polymerase chain reaction using standard protocols: 25-50 ng template, 5 pmoles each primer (Table 2.2), 1.25 nmoles each dNTP, 1X PCR Buffer (New England Biolabs, Ipswich, MA), 0.5 units Taq polymerase, and nuclease-free H_20 to 25 μl. Amplicons were purified with Exonuclease I

Table 2.1. Datasets from literature used to assess gene tree discordance at different depths of phylogeny.

Study	Taxon	T(mya)	Loci*
Jennings and Edwards (2005)	*Poephila* (sister species)	<1	25
Wiens et al. (2008)	Colubridae (family)	40	18
Wiens et al. (2008)	Serpentes (suborder)	90	15
Vidal and Hedges (2005)	Squamata (order)	160	9

*Number of loci used in this study; some loci data sets were incomplete and thus not used.

and Antarctic phosphatase following Glenn and Schable (2005). Fragments were sequenced following manufacturer's protocol, and sequences were analyzed on an ABI 3130 Sequence Analyzer (Applied Biosystems, Foster City, CA). When heterozygotes were detected, we first attempted to determine phase based on sample parameters using Phase (Stephens and Donnelly, 2003; Stephens et al., 2001). For those whose estimated phase had a posterior probability less than 0.95, amplicons were cloned using a QIAGEN cloning kit, and sequenced multiple clones per heterozygous individual to determine the exact phase.

Table 2.2. Primer pairs used in this study.

Primer	Gene	Oligo (5'-3')	Reference
BDNF-F	BDNF	GACCATCCTTTTCCTKACTATGGTTATTTCATACTT	Leache and McGuire (2006)
BDNF-R		CTATCTTCCCCTTTTAATGGTCAGTGTACAAAC	
FSHR_f1	FSHR	CCDGATGCCTTCAACCCVTGTGA	Wiens et al. (2008)
FSHR_r2		RCCRAAYTTRCTYAGYARRATGA	
Lglu	CYTB	TGATCTGAAAAACCACCGTTGTA	Alfaro and Arnold (2001)
H15544		AATGGGATTTTGTCAATGTCTGA	
G482	MC1R	TCAGCAACGTGGTGGA	Austin et al. (2009)
G480		ATGAGGTAGAGGCTGAAGTA	
ND4	ND4	TGACTACCAAAAGCTCATGTAGAAGC	Forstner et al. (1995)
M246		TTTTACTTGGATTTGCACCA	Skinner et al. (2006)
NTF3_F1	NT3	ATGTCCATCTTGTTTTATGTGATATTT	Wiens et al. (2008)
NTF3_R1		ACRAGTTTRTTGTTYTCTGAAGTC	
L75 F	R35	TCTAAGTGTGGATGATYTGAT	Fry et al. (2006)
H792 R		CATCATTGGRAGCCAAAGAA	

Gene tree estimation.—For each dataset, we generated a maximum likelihood estimate of genealogy for each nuclear gene and the concatenated mitochondrial data. After checking alignment by eye, DT-ModSel (Minin et al. 2003) was used to select the model of evolution that best fit each fragment, and a heuristic search was performed in PAUP* (Swofford, 2003) to estimate the ML tree. Support for each gene tree was assessed by performing 1000 heuristic search bootstrap replicates.

Concatenated phylogenetic analyses.— Phylogeny was estimated for *Nerodia* were using both a likelihood and Bayesian approach. A maximum likelihood phylogeny was estimated using PAUP*. The best model for the concatenated dataset was chosen using DT-ModSel (Minin et al., 2003), subsequently a heuristic search was performed using

estimated model parameters. Statistical support was assessed with 1000 heuristic search bootstrap replicates. Two Bayesian methods were also utilized to estimate phylogeny: MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) and BEAST (Drummond and Rambaut, 2007).

Species tree estimation.— The methods that currently exist for estimating species trees can be placed into two categories: those which estimate a species given estimated gene trees as input (eg., MDC [Maddison, 1997; Maddison and Maddison, 2009] and STEM [Kubatko et al., 2009]) and those which simultaneously estimate the gene trees and species tree (BEST [Liu et al., 2008], *BEAST [Heled and Drummond, 2010]). The former class relies on simple algorithms to estimate the species tree, whereas the latter uses Markov chains (one in *BEAST; multiple in BEST) to approximate the posterior probabilities of trees and parameters. These Bayesian methods are often computationally intensive, thus the "gene tree input" approaches may be preferred when no a priori reason for method choice exists. However, these methods rely on the assumption that the gene trees are well-estimated, which may not be the case in many empirical datasets, and inclusion of poor estimates of gene trees into studies may decrease the accuracy of species tree estimates. Empiricists are in a difficult position, as there is no simple measure of accuracy for gene trees estimated from empirical data because the actual genealogy is unknowable. We proceed here by estimating species trees using both approaches (i.e., species tree from gene trees and simultaneous estimation of species and gene trees) and conducting several simulation studies to enhance our understanding of how accurate we can expect various methods to be given our data.

Species tree from gene trees estimation.—Mesquite (Maddison and Maddison, 2009) was used to estimate the species tree by minimizing the number of deep coalescences (Maddison and Knowles, 2006). AUGIST (Oliver, 2008) was used to assess nodal support using trees saved from the non-parametric bootstrap analysis. Since Mesquite produces an estimate of the topology but not the branch lengths, STEM (Kubatko et al., 2009) was used to identify the ML estimate of the species tree (with branch lengths) given the gene trees. For both analyses, maximum likelihood estimates of the gene trees were used.

Bayesian species tree estimation.—Two methods for estimating species trees in a Bayesian framework are currently available, both of which simultaneously approximate the posterior distribution of the gene trees and the species tree, given multi-locus datasets and distributions of parameter priors. For Bayesian Estimation of Species Trees, BEST (Edwards et al., 2007; Liu, 2008), we conducted two runs of seven chains (one for each gene tree, species tree), for 10^8 generations, sampling every 10^4 generations. We used an inverse gamma distribution with shape parameters $\alpha = 3$, $\beta = 0.003$ ($\Theta = 0.0015$) for the theta prior and a uniform gamma prior with bounds 0 and 5, with the upper bound corresponding to k-1 independent loci (D. Rabosky, pers. comm.). Convergence of chains was assessed using the program AWTY (Wilgenbusch et al., 2004). We also used *BEAST (Heled and Drummond, 2009), which is implemented in the BEAST 1.5.2 software package (Drummond and Rambaut, 2007). *BEAST uses a single MCMC chain to estimate both the species tree and the gene trees; this chain was allowed to run for 10^9

generations, sampling every 10^5 generations. The first 2000 samples were discarded as burn-in, and each parameter was checked for autocorrelation using the program LogCombiner provided in the BEAST package. A maximum clade credibility tree was created using Tree Annotater, also provided with the BEAST package.

2.2.2 Simulations

Quantifying the lingering effects of coalescent variance.— To better understand how the coalescent processes that acted on the ancestral nodes of phylogenetic trees can influence phylogeny estimation, we conducted a series of analyses using data simulated in Mesquite 2.7.2 (Maddison and Maddison, 2009). For each of ten species topologies simulated under a birth/death process, we simulated 20 coalescent gene trees (20 alleles) contained within each species for four depths: 1N, 10N, 100N, 1000N. For each of these topologies, we made pairwise comparisons of topology (RF distances) using PAUP*. Then, for each gene tree at all depths, DNA sequence data was simulated using the average fragment lengths and models of evolution from the empirical datasets that most closely resemble the each species tree depth (Table 1). For these simulations, effective population size was set to N_e=10,000 and a generation time of 2.5 years was used (Gibbons and Dorcas, 2004); estimated node ages based on fossil data (Holman, 2000; Evans, 2003; Apesteguia and Zaher, 2006) were converted to *N* generations. We estimated a ML tree for each simulated dataset under the model of evolution used to simulate the data, and once again compared the topologies using RF distances. To measure how much phylogenetic estimation error affects the topology, comparisons of distributions of RF distances of the simulated gene trees and their respective estimated gene trees were performed. Finally, in order to discern the effect of gene tree discordance on phylogenetic inference, a concatenated estimate was produced for each species tree and compared to the simulated topology using RF distances and the metric employed by Kuhner and Felsenstein (1994), implemented in Ktreedist (Soria-Carrasco et al., 2007), which calculates relative Kuhner-Felsenstein (KF) distances for trees of differing total length. Because all fragments were simulated under the same model parameters, we did not partition the data for analysis.

Identifying the cause of gene tree incongruence.—For any two gene trees estimated from independent loci, some combination of phylogenetic estimation error and coalescent uncertainty can account for observed topological discordance. Determining the relative contributions of these processes is a vital step towards determining which of the competing paradigms to use to estimate the species tree. To test whether incongruence among gene trees could be attributed solely to phylogenetic estimation error, we use the parametric bootstraps (Swofford et al., 1996), an approach that utilizes simulation to construct a null distribution of the amount of phylogenetic error expected under a null model of no difference in topology across genes. We conducted pairwise test for all loci; in each we constrained the ML tree search of gene “A” to trees that matched the topology of gene “B”, then measured the deterioration of the likelihood score between the topologically unconstrained and constrained trees ($-lnL_{uncon} - -lnL_{con} = \delta lnL$) using PAUP. We then simulated 1000 datasets under the model and parameters of gene “A” on the topology of gene “B” using Seq-Gen (Rambaut and Grassly, 1997) and built a null

distribution of δlnL to examine our test statistic. Since parametric bootstrapping depends on an adequate fit of the model of sequence evolution to the data (Goldman et al., 1996), we conducted an absolute goodness-of-fit tests on each gene and corresponding model with a modified method of Sullivan et al. (2000), using 1000 simulated datasets. For both tests, significance was assessed using a Bonferroni corrected α ($0.05/n$ comparisons).

Gene tree support and species tree accuracy.—The quality of a maximum likelihood phylogenetic estimate is typically assessed by calculating non-parametric bootstrap for each node of the phylogeny (Felsenstein, 1985). To test if our set of gene trees estimates (with assessed nodal support) were sufficiently accurate for STEM to recover the correct species tree, a series of simulations were conducted. Starting with the topology of the species tree estimated from the empirical data using *BEAST, we simulated 1000 coalescent gene trees, consisting of two alleles per species with an N_e of 10,000 and a total tree depth of 50N, using Mesquite. For each gene tree, sequence data was simulated using the program Seq-Gen (Rambaut and Grassly, 1997), under the HKY model of sequence evolution, with nucleotide frequencies and length (551 bp) of each fragment taken from a mean of the empirical data. Each dataset was simulated on a tree with a length drawn from an exponential distribution with mean 0.05 substitutions/site (i.e., the mean tree length of the nuclear gene tree estimates). Maximum likelihood gene trees were then estimated under the same model under which they were simulated, and 100 "fastsearch" bootstrap replicates was performed for each tree. Gene tree quality was assessed in two ways. First, a measure of average nodal support (ANS) was calculated for each tree as the sum of all nodal support values above 50 divided by the total number of potentially supported nodes (18).

We used STEM to estimate a species tree from 1000 subsets of 6 randomly chosen ML gene trees; then the Kuhner-Felsenstein and Robinson-Foulds metric was calculated between the estimated and actual species tree to assess accuracy. This number was the compared with linear regression to both the mean and variance of ANS. Perl scripts were written to automate these simulations and are available on the senior authors web site.

2.3 RESULTS

Data collection and gene tree estimation.—A total of 3857 bp of phased DNA sequence data was collected for 13 individuals representing 10 species. Gene tree estimates for each gene are shown in Figure S1. Average nodal support across all gene trees as 47.6. This number is proportional to the number of segregating sites in each gene (data not shown). Descriptive statistics for each gene and model of evolution selected can be seen in Table 2.3.

Quantifying the lingering effects of coalescent variance.— For simulated species trees, coalescent gene trees showed some level of discordance at all depths (Fig. 2), in 9/10 and 1/10 topologies at 100N and 1000N respectively, while there was some discordance among estimated gene trees in all cases. Comparisons of RF distributions of actual and estimated gene trees indicate that in most cases, the primary source of topological

incongruence is phylogenetic error. In one of ten 100N and three of ten 1000N trees, concatenated ML estimates of the species tree differed from their respective simulated topologies.

Table 2.3. Descriptive statistics of sequenced loci.

gene	frag. length	var. sites	model implem.
BDNF	572	18	K2P
FSHR	512	25	K81uf + G
mtDNA	1133	319	GTR + G
MC1R	435	32	HKY+I
NT3	561	38	K2P + I
R35	659	28	HKY + G
Total	3857	460	N/A

Identifying the cause of gene tree incongruence.—Discordance between topologies was significant ($p < 0.002$) in 18 of 25 pairwise tests (Table 2.4; 23/25 were "significant" prior to correction for multiple comparisons). Comparisons testing fit of the FSHR gene to other topologies were not conducted, as the model was a poor ($p < 0.001$) fit to the data.

Gene tree support and species tree accuracy.—Results of this simulation exercise (Figure 2.3) are consistent with the prediction that accuracy of species tree estimation is directly correlated with quality of gene tree estimation. Average nodal support for the empirical dataset was 47.6, which was below the lowest simulated ANS for which the gene tree subset yielded the correct topology. Based on these results, results of gene tree-based species tree estimators are not presented.

Species tree estimation.—The maximum clade credibility tree obtained from *BEAST can be seen in figure 4. After 10^9 generations, effective sample sizes of all parameters were greater than 200 (the minimum suggested by the authors for publication). BEST results are not shown. After 10^8 generations, standard deviation of split frequencies for all gene tree chains were above 0.07; stationarity is assumed when these values are below 0.01. Convergence was not assessed as stationarity had not been reached**.**

2.4. DISCUSSION

We provide a simple framework which will allow researchers to make an a priori decisions about which model of phylogeny is best to use given their data: a simpler model in which all genes share a topologically identical history, or more complex models which allow genealogies to vary due to coalescent processes. In the first step, we compare the topologies of gene trees using a parametric approach. If topologies are not significantly different we could safely estimate our species tree using a concatenation approach, and additional loci can be gathered at the expense of within-species sampling.

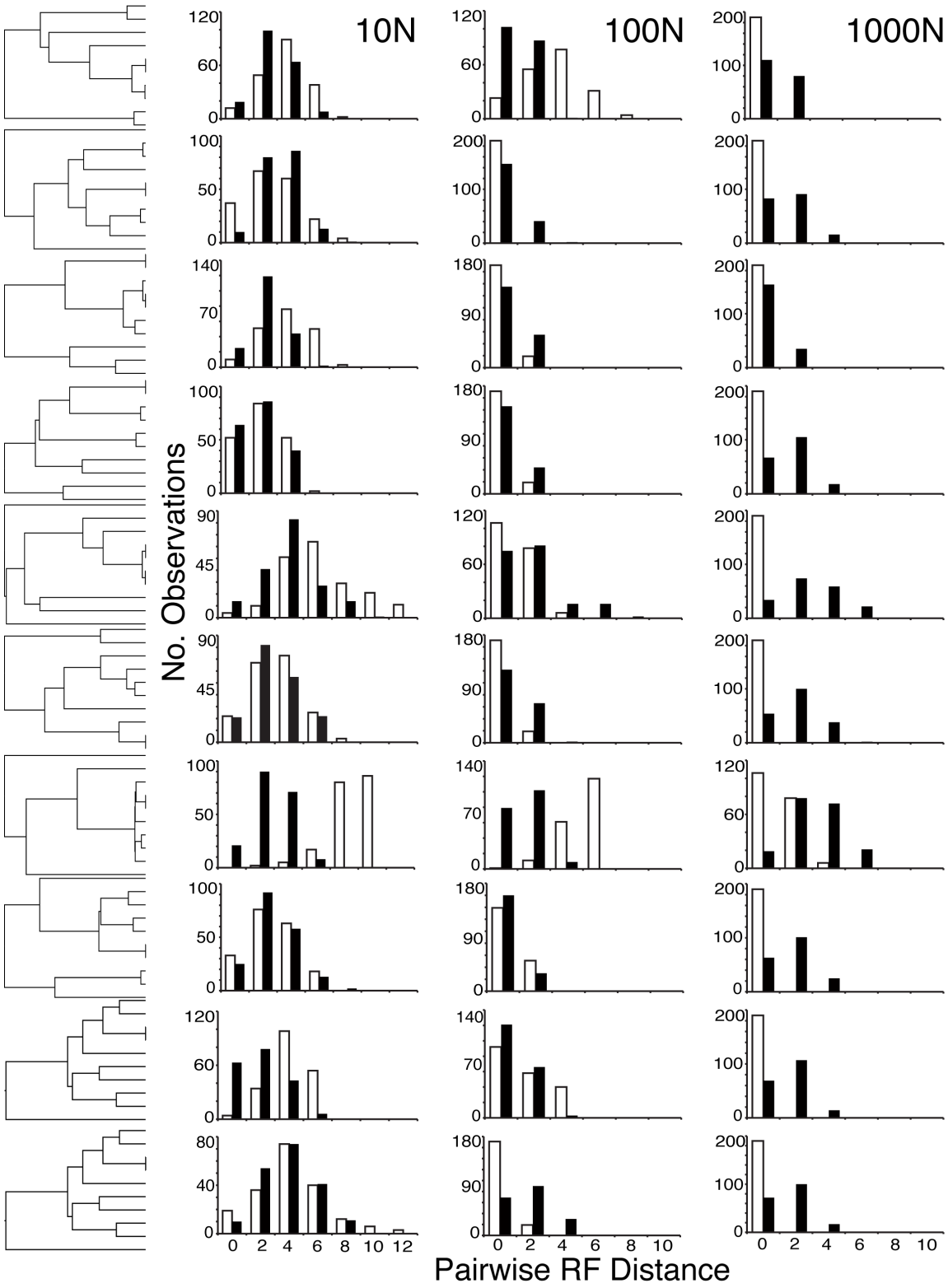


Figure 2.2. Distributions of Robinson-Foulds distances of actual (white) and estimated (black) gene trees. Trees on left indicate actual species tree under which coalescent genealogies were simulated.

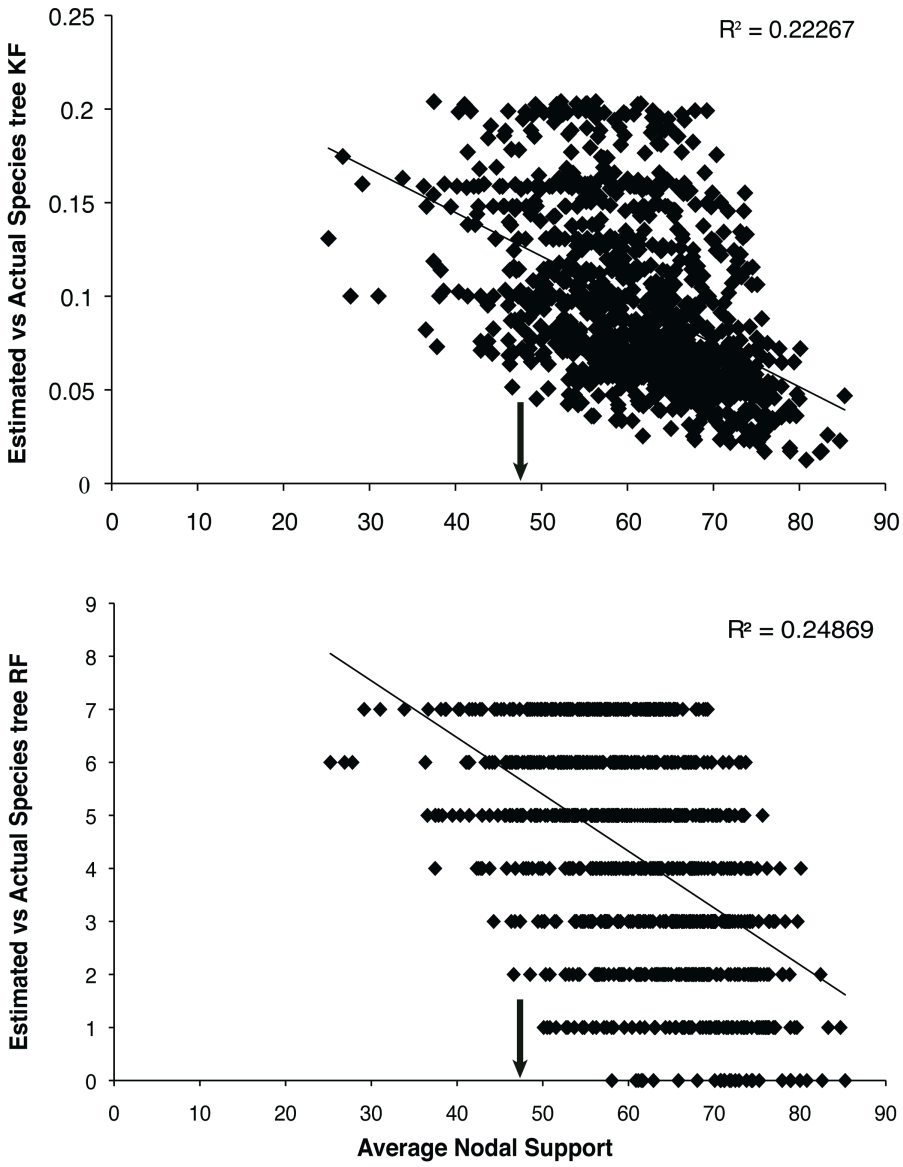


Figure 2.3. Comparison of average nodal support across a subset of gene trees and their utility in species tree estimation, using a) Kuhner-Felsenstein distances and b) Robinson-Foulds distances. Arrow indicates average nodal support of empirical dataset.

However, if gene trees exhibit an amount of incongruence that can not be attributed to phylogenetic estimation error alone, then a coalescent-based approach is preferred. For our data this is clearly the case as some 18/25 of our comparisons were able to reject the null hypothesis (i.e., that there is no difference in the topology of gene A and gene B) even using the conservative Bonferroni correction. Faced with these results, we attempted to determine if our gene trees were estimated sufficiently well to produce accurate results using STEM, since this program produces accurate estimates of species phylogenies when the gene trees are estimated without error (Kubatko et al. 2009; McCormack et al.

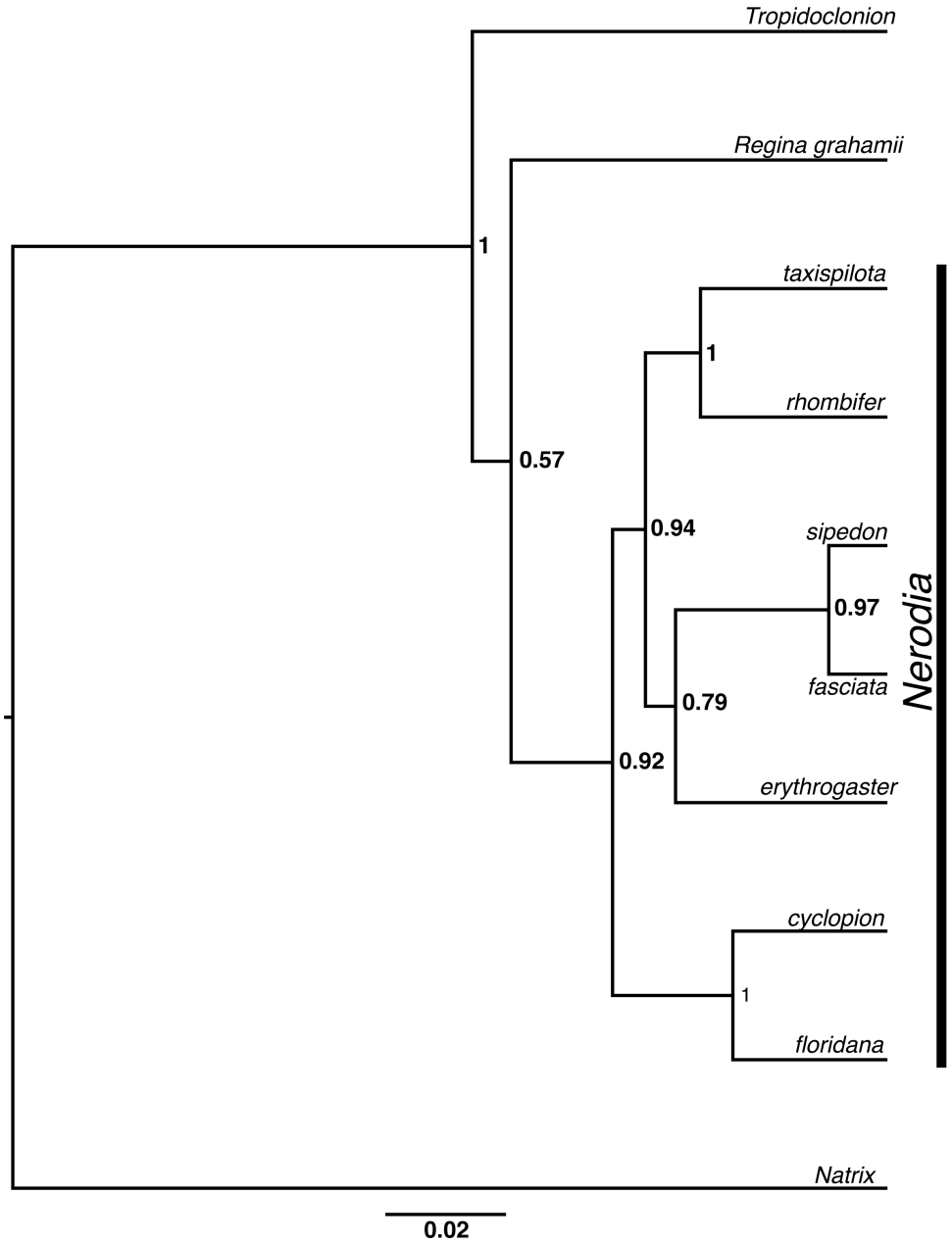


Figure 2.4. Maximum clade credibility tree based on 5 nuclear and one mitochondrial locus from *BEAST.

2010). We proposed a procedure based on the calculation of the average nodal support; trees that are estimated with little error will tend to have highly supported nodes as measured by non-parametric bootstrapping. Our results indicate that the ANS is low for our system, suggesting to us that we can not be sure of the accuracy of the species tree estimate from STEM. Therefore, we simultaneously estimated the posterior distributions of our gene trees and species tree using *BEAST.

Relationships among *Nerodia*.—Results of the *BEAST analysis recovers *Nerodia* as a monophyletic clade, with reasonable support at most nodes. Disagreement between our

estimation and that of Alfaro (2003) may be due to several factors (e.g., taxon sampling in terms of the species representing the ingroup, the total number of individuals per species, and the total number of taxa included in the analysis). Results of our analysis will likely change as taxa and individuals are added, as our ultimate goal is to estimate phylogeny of Thamnophiini. Given our findings from these preliminary data; we are presently collecting data from multiple individuals in all (~60) of the Thamnophiini species and will present a more densely-sampled and rigorous phylogeny as Chapter 2. However, our results illustrate several striking patterns that have clear implications for empirical systematists.

Quantifying the lingering effects of coalescent variance and phylogenetic estimation error.— Our first set of simulations suggest that anomalous lineage sorting due to coalescent stochasticity can result in gene tree discordance even in phylogenies with large total tree depths. This is perhaps not surprising, since discordance should be expected within any species trees that possess internode less than ~6N generations in length (Hudson and Coyne, 2002) somewhere within the tree. However, we determined that, at deeper phylogenetic levels, phylogenetic estimation error was a more common source of estimation error than coalescent uncertainty. While these would seem to imply that concatenation would perform well, it is generally the case that *both* of these processes contribute to decreased accuracy in phylogeny estimation. We advocate parametric bootstrapping as a method for determining whether the observed incongruence across multiple loci can be attributed to phylogenetic estimation error alone.

Gene tree support and species tree accuracy.—One of the approaches used by empiricists to measure the quality of their phylogeny estimates is the bootstrap support of particular nodes in the phylogeny. We extend this convention and measure the overall quality of our gene tree estimates by averaging the nodal support, and then used regression to demonstrate that the accuracy of species trees estimated using STEM are correlated to the ANS. Based on results of the third set of simulations, we consider the information contained within our gene trees inadequate to estimate gene genealogies sufficiently for use in gene tree-based species tree estimators. Species tree estimates using STEM and Mesquite differed in topology from both the concatenated and *BEAST results, with STEM not recovering *Nerodia* as a monophyletic group (Appendix A, Fig. S3). It is possible that these estimates will improve when numbers of alleles per species are increased (Hird et al., in review).

Causes of discordance.—While we have evoked the explanation for discordance among our gene trees as coalescent stochasticity, there are other phenomena, such as hybridization and gene duplication/extinction, which may cause similar patterns in the observed data. While a theoretical and practical framework are currently being developed that incorporate hybridization into coalescent-based analyses of phylogenetic estimation (Kubatko, 2009; Meng and Kubatko, 2009), this long-standing problem remains a difficult one. A key to useful incorporation of hybrid mechanisms may be a better understanding of the system-specific patterns of introgression. Within Thamnophiini, hybridization has been shown to occur between both sister and non-sister pairs of species

(Fitzpatrick et al., 2008; Mebert, 2008); consequently we cannot ignore hybridization as a possible mechanism influencing the patterns we observe.

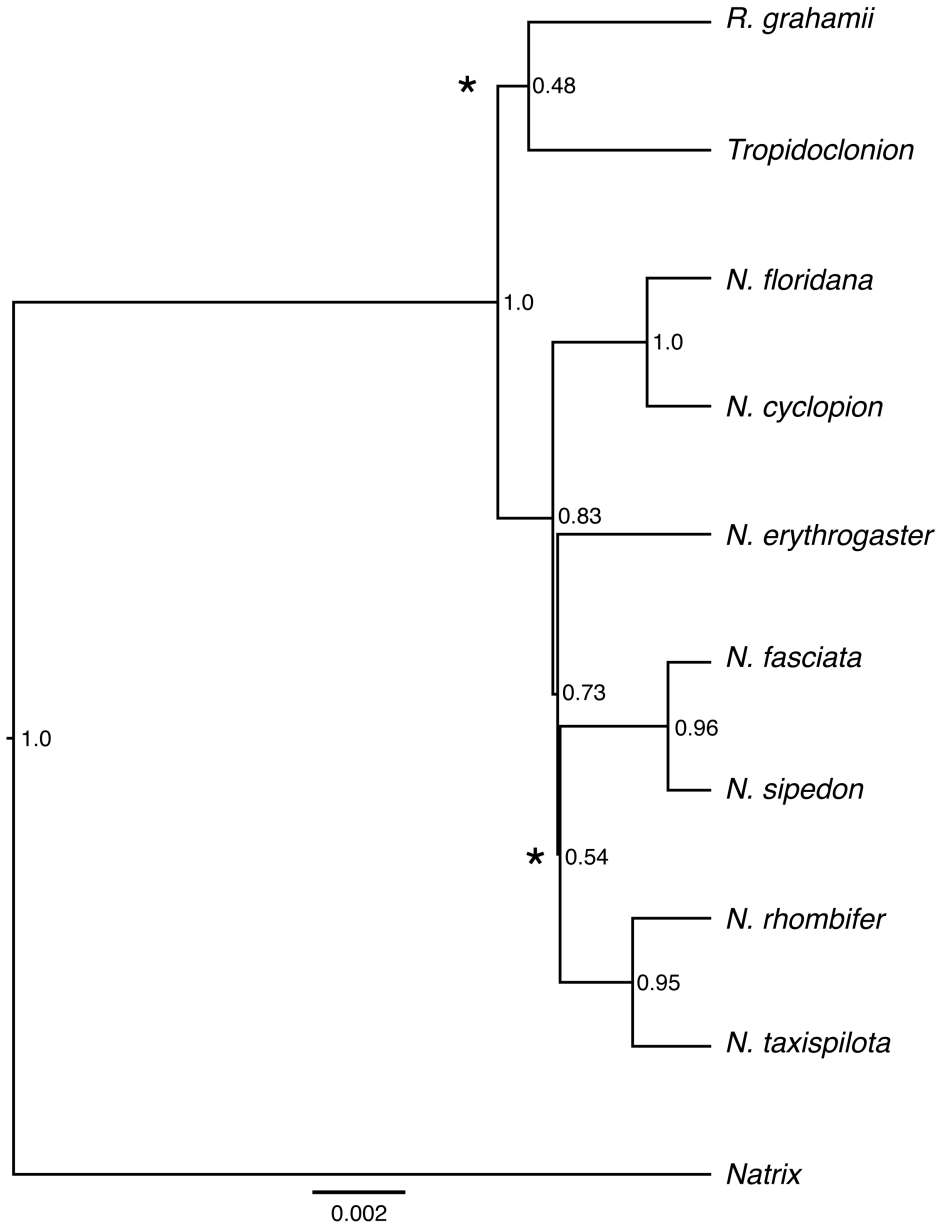


Figure 2.5. Maximum clade credibility tree based on 5 nuclear from *BEAST. Nodes at which there is disagreement between estimates which include and exclude mitochondrial data is denoted by *.

Rate heterogeneity and species tree estimation.—An advantage to using species tree estimators that require gene trees as input is that each gene tree contributes equally to the likelihood of the species tree, thus no single gene tree topology can disproportionally influence the species tree estimation. This is not the case with concatenation. Disconcertingly, concatenated multilocus phylogenetic estimations often include one or more mitochondrial loci, and the sheer bias in the number of variable sites is likely to

result in an estimation in which the signal from the nuclear data is treated a noise, overwhelmed by the information in the plasmid loci (Carstens and Knowles, 2007b). Even if the mitochondrial genealogy is concordant with other gene trees or the species phylogeny, the species tree estimate could still suffer from a bias in branch length estimates, which can result in incorrectly estimated node ages, or bias in ancestral character state reconstruction. However, we suggest below that species tree estimates may be subject to similar biases caused by dramatic differences in the amount of phylogenetic information across loci.

Our data include both mitochondrial and nuclear loci; it happens that the topology of the tree obtained from *BEAST is similar in topology to that of the concatenated estimate, particularly within *Nerodia* (Appendix A, Fig.S1) and also to the gene tree estimated from the mitochondrial data alone (Appendix A, Fig. S2d). Since the Markov chain samples the posterior distribution of tree space using all the data, we suspect that the sampling of species tree topology may be biased in favor of concordance with the topologies of the gene trees that contain the most phylogenetic information (here, the mitochondrial data). To explore the possible bias in our species tree estimate, we performed another *BEAST analysis, under previous conditions, but did not include the mitochondrial data. Topologies differed among the two trees (Fig. 5), however support for the relationships was low in the latter estimate. Simulations studies to explore the possible influence of a variable-site rich mitochondrial locus on a simultaneous gene tree/species tree estimate are needed.

Conclusions.—We demonstrate a useful and direct approach to choosing among the two dominant phylogenetic models; concatenation and species tree estimation. Central to our description of the issues related to choosing among these models is the assumption that it is important to have an *a priori* expectation of model performance in order to avoid a *post hoc* evaluation of the phylogenies. We also contend that the optimal sampling design differs for these competing models; for our data we show clearly that coalescent processes are likely to produce incongruence across loci and therefore future efforts will be focused on increasing the number of individuals included in the analysis.

CHAPTER 3.
TESTING MONOPHYLY WITHOUT WELL-SUPPORTED GENE TREES: MULTI-LOCUS NUCLEAR DATA CONFLICT WITH EXISTING TAXONOMY IN THE SNAKE TRIBE THAMNOPHIINI

3.1. INTRODUCTION

Accurate estimates of phylogeny from molecular data offer vital information for understanding the evolution of any clade of organisms, particularly those that exhibit apparent lability in key morphological traits. One such group is the Natricine snakes; these snakes occur on all continents but Antarctica and S. America, and are represented in the New World by approximately 60 species (tribe Thamnophiini). Natricine snakes are a particularly compelling focus for phylogenetic analysis because representatives of this group occupy broad ecological niche space, from complete terrestrial to almost exclusive aquatic habitat, with a corresponding breadth of feeding niches ranging from broad generalists to stenophagic diets (Gibbons and Dorcas, 2004). Data from molecular phylogenies enable us to understand the evolutionary lability of traits related to habitat and diet, and to estimate how quickly these traits evolve.

Members of the tribe Thamnophiini are particularly flexible in diet, with members feeding on fish, amphibians, reptiles, insects, mollusks, and crustaceans. One feeding specialty highlights the important role played by molecular phylogenies: the four crayfish specialists in the genus *Regina*. There is debate among scholars as to whether these snakes represent a single or two monophyletic groups. While ecology and feeding behavior suggest a shared ancestry, other characters such as dental morphology (Rossman, 1963), and scale microtexture (Price, 1983) have led some scholars to place two of these species, the Glossy (*R. rigida*) and Striped Crayfish Snake (*R. alleni*) into their own previously described genus, *Liodytes* (Cope, 1885); but see Rossman (1963, 1985). Most recently, Alfaro and Arnold (2001) suggested that these species do not represent a monophyletic lineage, based on phylogenetic analysis of mitochondrial sequence data; however they made no specific taxonomic recommendations because their results were not sufficient to make robust conclusions. Interestingly, one well-supported clade from this contained the two "*Liodytes*" species as well as the Swamp Snake (*Seminatrix pygaea*), which lacks many of the derived morphological characteristics present in the other two species. If this phylogeny is accurate, the feeding specializations associated with crayfish eating are either convergent or have been lost in some members. However, this inference depends on the accuracy of the phylogeny estimate; in this case the phylogeny was produced from three mitochondrial genes (ND2, CYTB, 12S) for 27 ingroup species representing eight genera (Alfaro and Arnold 2001).

The Thamnophiini also contain the earth snakes (genus *Virginia*), another problematic group represented by two species (*V. striatula* and *V. valeriae*). Originally placed in the novel genera (*Haldea striatula* and *V. valeriae*) by Baird and Girard (1853), the former was submerged within the *Virginia* by Garman (1883), and has since had a number of studies supporting or rejecting this move (for a review, see Rossman and Wallach [1991]), including allozyme data that lends support to original designation, however no

taxonomic changes have been formally accepted. To date DNA sequence data has only been published for one species, leaving this taxonomic change untested in a modern phylogenetic framework.

Here we use molecular sequence data to address the taxonomic status of the two natricine snake genera *Regina* and *Virginia*, the latter of which contains two (or three, according to some authors (Collins and Taggart, 2002)) species, but to date has only been represented by one species in molecular genetic studies. Specifically, we ask "Is *Regina* a monophyletic genus, or does it represent two or more independently evolving lineages?" and "Are the earth snakes (genus *Virginia*) sister taxa?" We will address these questions with a multi-locus, mitochondrial and nuclear dataset containing one or more representatives of all putative genera in Thamnophiini.

3.2. METHODS

3.2.1. Sample preparation and sequencing

Tissue samples of specimens were obtained from the Louisiana State University Museum of Natural Science (Appendix A Table 1). We extracted total DNA from tissues following a modified version of the protocol described by Aljanabi and Martinez (1997), where tissues were initially digested overnight in 300µL of Puregene® Cell Lysis Solution (QIAGEN catalog no. 158906) and 2.5µL Proteinase K (New England Biolabs no. P8102S) prior to following the standard protocol. DNA samples were then quantified via Nanodrop (Thermo Scientific, Waltham, MA) and diluted to a final concentration of 10-25ng/µL.

Polymerase chain reactions were performed for each individual for five nuclear and two mitochondrial genes (Appendix A Table 2). Reactions were performed in 25µL with the following reagent concentrations: 0.4-1ng/µL tDNA, 0.4µM each primer, 0.2µM dNTPs, 1X Standard Taq reaction buffer (New England Biolabs) and 0.5 units of Taq DNA polymerase (New England Biolabs no. M0267). For all but ND4 (55°C annealing temperature), thermocycling was performed with an initial melting step of 2 minutes at 95°C 30 cycles of: 30 seconds at 95°C, 15 seconds at 50°C and 30 seconds at 72°C, followed by 10 minutes at 72°C. Sequence analysis was performed on an ABI 3130XL (Applied Biosystems, Foster City, CA) after sequencing was performed using BigDye v 3.1, following manufacturers instructions. Both PCR and cycle sequencing products were purified following an ethanol precipitation procedure. Sequences were edited using Sequencher 4.8 (Genecodes, Ann Arbor, MI) and aligned using Muscle (Edgar, 2004) and manually verified. Phasing of ambiguous alleles was performed using PHASE 1.4 (Stephens et al., 2001); data was formatted using SEQPHASE (Flot, 2010). See below for treatment of sites that could not be resolved with greater than 90% confidence using PHASE.

3.2.2. Gene tree estimation

Phylogenetic estimates were produced for each nuclear gene fragment and the combined mitochondrial fragments using MrBayes 3.2.1 (Ronquist et al., 2012). We chose models of sequence evolution following results from DT-ModSel (Minin et al., 2003). For each gene, a four-chain (three cold, one hot) Markov Chain Monte Carlo (mcmc) was run for 5,000,000 generations, sampling every 500, or until standard deviations of split frequencies fell below 0.01, ensuring proper mixing of chains. A burn-in of 25% of sampled steps (program default) was used for all genes; support for nodes was assessed using Bayesian posterior probability (BPP) values.

An important assumption of the coalescent model is that genes are evolving in a neutral fashion. Violation of this assumption may lead to branch length heterogeneity among gene trees (Edwards, 2009) and instances of strong directional selection may lead to topological bias in the estimated gene tree. After determining reading direction and frame for each gene, we tested for evidence of purifying selection (dS>dN) by using the codon-based z-test of selection in Mega5 (Tamura et al., 2011), implementing the overall average function.

3.2.3. Phylogeny estimation

Because any recombinant unit of DNA (such as the mitochondrial genome) is subject to the stochastic process of gene coalescence, its genealogy may not reflect the actual pattern of species divergence (Degnan and Rosenberg, 2009). Thus, any phylogenetic estimate based on sequence from a single recombining unit may be biased in both branch length and topology, and assuming a single underlying genealogy for multiple, unlinked genes (as is the case when data are concatenated) may lead to biased or even positively misleading estimates (Degnan and Rosenberg, 2006) of the containing phylogeny.

Estimates of phylogeny under both the concatenated and multi-species coalescent model were produced using BEAST 1.7.1 (Drummond et al., 2012). Substitution models obtained from DT-ModSel were implemented for each gene, substitution rates and clocks were unlinked across genes, and clock model for each gene was set to uncorrelated relaxed – lognormal, with uniform prior with a range of 0-10. The MCMC was run for 100,000,000 generations, sampling every 10,000 generations. For the multi-species coalescent, topologies were unlinked among genes (mitochondrial topologies remained linked) and the *BEAST prior was implemented. Each analysis was performed twice, and posterior distributions of parameters were compared to ensure consistency across runs using Tracer 1.5 (Rambaut and Drummond, 2009).

3.2.4. Phasing

Coalescent-based species tree estimators rely on population genetic parameters and processes to estimate relationships among populations. Parameters like $\theta = 4N_e\mu$ can be better estimated given more information about the allele frequencies within populations. When more than one site is ambiguous within an individual sequence, determining the

alleles (experimentally or via estimation) representing this sequence is important for estimating coalescent parameters, and can reveal anomalous shared ancestry of alleles among populations. As the populations being studied become more distantly related, the probability of shared alleles becomes lower, leading to the idea that phasing of ambiguous data will be less important to estimation of gene trees and the containing species tree. Thus, in a case such as ours, where divergence between the species included in the investigation are likely greater than the expected time to coalescence of all alleles within a given species, ambiguous sites that cannot be phased may represent allelic autapomorphies that will not affect the outcome of the analysis. To test whether this is the case, we used Paup* (Swofford, 2003) to build neighbor-joining trees containing all possible phases for each gene with ambiguous data, with the null expectation that all possible alleles should form a monophyletic group. Depending on the depth of relationships being investigated, violation of this expectation may be attributed to one or more causes, including incomplete or anomalous lineage sorting, introgression, gene duplication and loss, and selection.

3.2.5. Tests of monophyly

Monophyly of a previously designated or putative taxonomic group may be rejected when there is statistical support based on BPP for a branch or branches within trees that contain topologies incongruent with the taxonomic hypothesis. However, when monophyly is not supported by the phylogenetic estimate, but there is not enough statistical support (i.e. BPP >0.95) to reject monophyly, a comparison of marginal likelihood estimates between two models, one topologically constrained to include the monophyletic clade to be tested, and one topologically unconstrained. Here we test a number of putative monophyletic groups within Thamnophiini: 1) *Regina* (Rossman, 1963) 2) *R. septemvittata* and *R. grahami* (to the exclusion of the *Liodytes* group; Lawson [1985]) 3) *Liodytes* (Price, 1983), 4) *Liodytes* and *Seminatrix* (Alfaro and Arnold, 2001) and 5) *Virginia* (Garman, 1883). We constrained the topology for each of the above groups and performed a stepping stone run of 10,000,000 generations (50 steps with stationarity being reached in each step) from which we obtained a marginal likelihood estimate; each of these estimates were compared using Bayes Factors to the marginal likelihood estimate obtained from an unconstrained MrBayes run of the same length. The stepping stone function (Xie et al., 2011) implemented in MrBayes 3.2, offers an improved estimation of marginal likelihood over harmonic mean estimation. In addition to these tests, we measure support for each of the above groupings by observing both BPP >0.95 that support or exclude monophyly; we additionally filtered and counted trees containing these groups for each gene tree and species tree posterior distributions using the constraint filter commands in PAUP*.

3.3. RESULTS

3.3.1 Gene trees

A variation of the two-substitution site model was chosen for each gene (Table 1). Within each gene, topologies of consensus trees (maximum clade credibility) were consistent

between runs. Support (BPP) was generally low among all nuclear loci, but high for many nodes within the mitochondrial gene tree estimate (Fig. S1). Topologies were inconsistent among genes; however no strongly supported discordance between topologies was present. The codon-based z-test of selection strongly rejected ($p<0.01$) neutrality in favor of purifying selection across all genes tested (Table 1).

Table 3.1. Bayes Factors of stepping-stone-based estimates of marginal likelihood for five putative monophyletic groupings. Strong favor (-5 to -3), substantial favor (-1.5 to -3), substantial rejection (1.5-3), strong rejection (3-5), very strong rejection (5-6.6), decisive rejection (>6.6).

		Gene					
Taxonomic Constraint		BDNF	FSHR	MC1R	MT	NT3	R35
	+	-1082.39	-1057.59	-1013.34	-6399.34	-1441.52	-1454.82
"*Regina*"	-	-1079.03	-1054.5	-1003.93	-6338.73	-1432.61	-1441.88
	BF	**3.36**	**3.09**	**9.41**	**60.61**	**8.91**	**12.94**
	+	-1079.22	-1058.33	-1002.63	-6341.72	-1434.54	-1445.69
"*Virginia*"	-	-1079.21	-1054.47	-1003.91	-6338.55	-1432.6	-1441.86
	BF	**0.01**	**3.86**	**-1.28**	**3.17**	**1.94**	**3.83**
	+	-1078.65	-1053.27	-1009.79	-6347.99	-1430.45	-1440.63
"*Liodytes*"	-	-1079.09	-1054.6	-1003.91	-6338.76	-1432.84	-1442.23
	BF	**-0.44**	**-1.33**	**5.88**	**9.23**	**-2.39**	**-1.6**
Liodytes	+	-1078.53	-1051.81	-1014.88	-6336.13	-1429.03	-1439.02
+ *Seminatrix*	-	-1079.07	-1054.31	-1003.93	-6338.63	-1432.56	-1442.16
	BF	**-0.54**	**-2.5**	**10.95**	**-2.5**	**-3.53**	**-3.14**
R. grahamii	+	-1083.1	-1059.13	-1005.02	-6345.86	-1431.75	-1448.41
+ *R. septemvittata*	-	-1078.97	-1054.34	-1003.95	-6338.75	-1432.58	-1441.85
	BF	**4.13**	**4.79**	**1.07**	**7.11**	**-0.83**	**6.56**

3.3.2. Phylogeny estimation

Phylogenies were estimated under two evolutionary models: concatenation and a coalescent-based species tree approach (Fig. 1). Topologies were incongruent at multiple nodes across the tree, however posterior probabilities were low for discordant nodes. Both estimates split the ingroup into a clade consisting of mostly fossorial snakes (*Clonophis, Regina alleni* and *rigida, Seminatrix, Storeria and Virginia*) and a mostly semiaquatic and terrestrial group (*Adelophis, Nerodia, Regina grahamii* and *septemvittata, Thamnophis, and Tropidoclonion*). Neither estimate recovered *Regina* or *Virginia* as monophyletic, and the concatenated estimated rejected these groupings with greater than or equal to 0.95 Bayesian posterior probability.

3.3.3. Phasing

We used PHASE to estimate alleles for each nuclear gene; in all but two (MC1R, NTF3), alleles could be resolved with high confidence (Appendix B Fig. 2). For NTF3, all possible phases for each species coalesced prior to the nearest interspecific node (i.e., the possible alleles were monophyletic). For MC1R, alleles representing *Adelophis foxi* were not monophyletic: two possible phase resolutions yielded similar results (Appendix B Fig. 2). These heterozygous sites were excluded from analyses.

3.3.4. Tests of monophyly

We compared the marginal likelihood estimates of a topologically constrained and unconstrained run of MrBayes for five possible monohyletic groups within Thamnophiini (Table 3.1). Results strongly suggest that the classic taxonomic groupings of the crayfish snakes and the earth snakes are not valid, and there was also evidence against a monophyletic group containing *Regina grahamii* and *R. Septemvittata*. The only strong positive evidence is shown for the group containing *R. alleni* and *R. rigida* ("*Liodytes*"), along with *Seminatrix pygaea*, with mixed evidence for grouping the two "*Liodytes*" as sister.

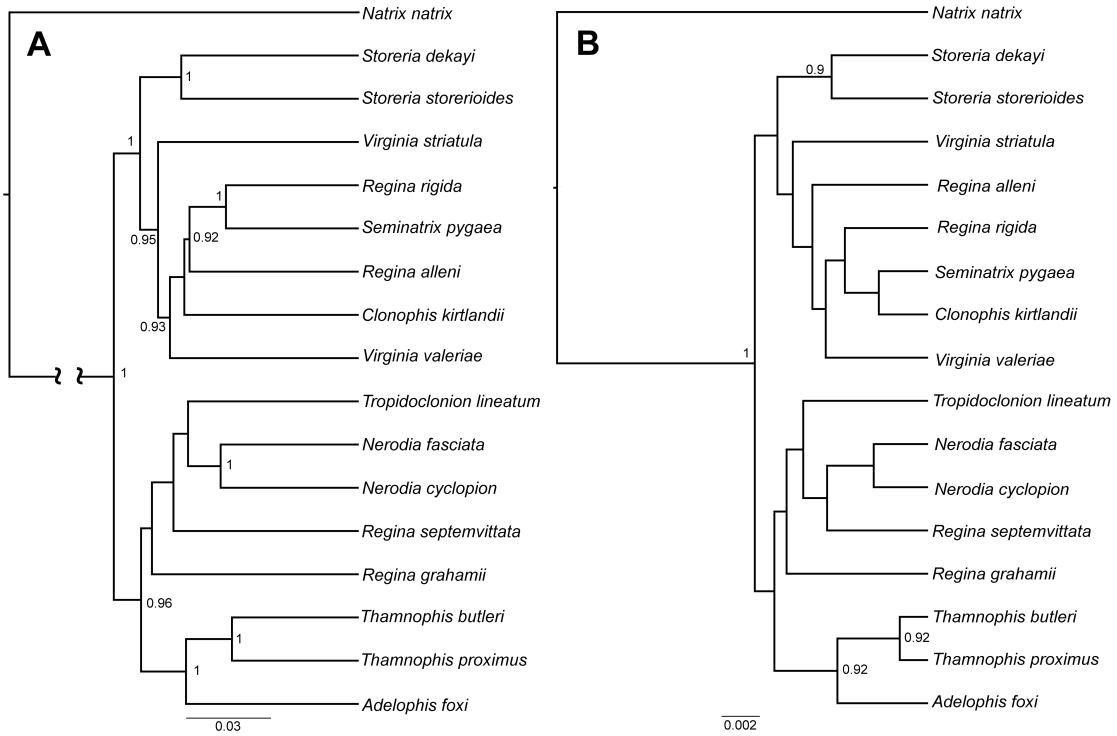


Figure 3.1. Multi-locus Bayesian maximum clade credibility estimates. A) Concatenated phylogeny; B) Multi-species coalescent. Unlabeled nodes were not supported with greater than 0.9 Bayesian posterior probability.

3.4. DISCUSSION

Taxonomists have traditionally employed morphological, ecological and distributional measures to diagnose and infer relationships among species. In the last several decades, molecular sequence data has played an increasing role in this field, and advances in both technology and methodology have led to changes in the way species are discovered and diagnosed (Wiens, 2007). However, molecular data remain one of several sources of data available to taxonomists, and methods for estimating phylogenies from these data have continued to evolve. It has been argued that the field of molecular systematics has been subject to a paradigm shift (Edwards 2009) related to with regards to how multilocus data are analyzed. Since we seek to recover the pattern of diversification across species, rather than to estimate a genealogy of a particular gene with the hopes that this genealogy reflects the underlying species tree, we favor species tree approaches that that model the

divergence of evolutionary lineages. We agree with Edwards (2009) that concatenation is not appropriate for the data collected here. However, we have also uncovered evidence that the data collected here are subject to purifying selection, and these results demonstrate that our data violate one assumption inherent to the species tree model (i.e., the coalescent model assumes selective neutrality).

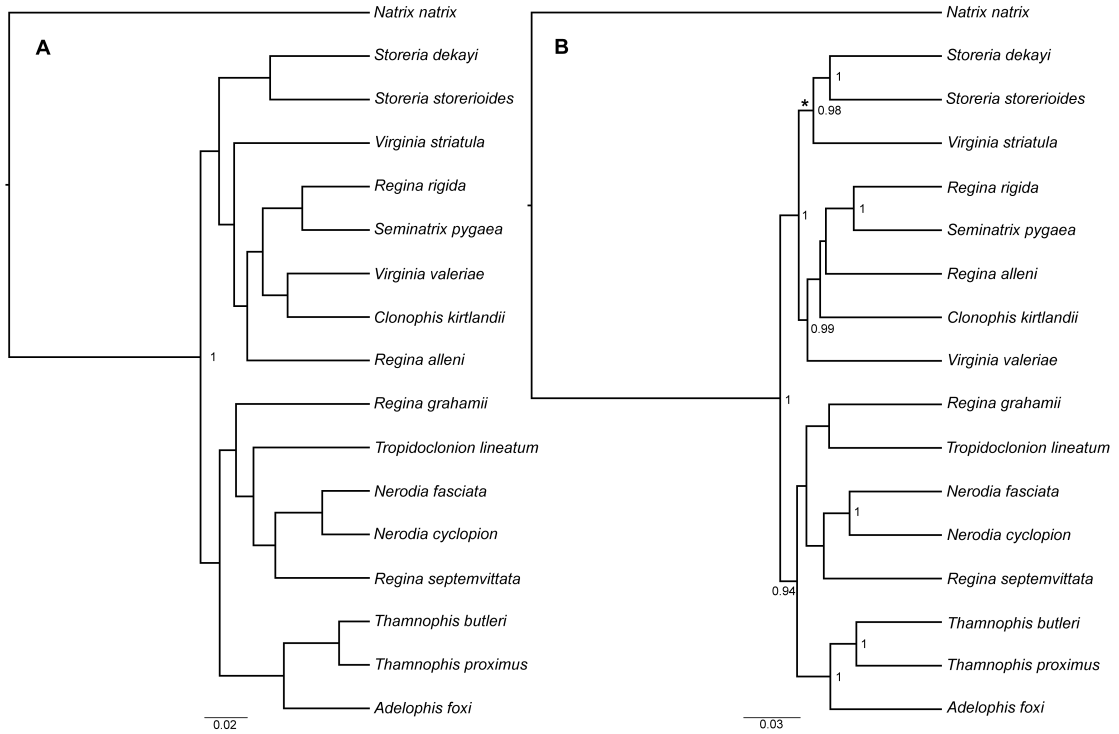


Figure 3.2. Multi-locus Bayesian maximum clade credibility estimates with MC1R data excluded. A) Multi-species coalescent; B) Concatenated phylogeny. *Indicates node in conflict (BPP >0.95) with analysis including MC1R

3.4.1 Tests of monophyly

Since neither concatenation or species tree analyses appear completely appropriate for our data, we quantified the support in the data for the taxonomic hypotheses on a gene by gene basis. We employed two techniques: filtering posterior topologies for trees containing groups in focus, and comparing the marginal likelihood estimates of positively and negatively constrained topologies using Bayes Factors. While we drew no strong conclusions from filtering the gene tree topology posteriors (Table 3.2), the results from the Bayes Factor-based tests of monophyly are clear in their interpretation. The relative power of the Bayes Factors is correlated with the number of segregating sites. On a locus by locus basis, we find little support for the taxonomic groupings of the earth snakes and crayfish snakes. Rather, our data follow that of Alfaro and Arnold (2001) in suggesting that these groups are unnatural paraphyletic (in the case of *Virginia*) or polyphyletic (in *Regina*) assemblages. In addition, there was a strong conflict in the measurable support for the *Liodytes* and *Seminatrix* clade, with four of six genes supporting the relationship and one (MC1R) rejecting it.

Table 3.2. Lines of evidence supporting or rejecting the five tested monophyletic groupings. Proportions of posteriors are the ranges of the proportions of distribution of topologies among gene trees estimated.

Taxonomic Grouping	Species Tree	Concatenation	Gene Tree Support	Proportions of Posteriors	BF Constraint Tests
"Earth Snakes"	No Support	Reject	2 Reject, 0 Support	0-0.03	4 Reject, 0 Support
"Crayfish Snakes"	Reject	Reject	3 Reject, 0 Support	0 in all genes	5 Reject
Liodytes	Reject	Reject	1 Reject, 0 Support	0-0.11	2 Reject, 2 Support
Liodytes + *Seminatrix*	No Support	Support	0 Reject, 0 Support	0.007-0.13	1 Reject, 4 Support
R. grahamii + *R. septemvittata*	No Support	No support	0 Reject, 0 Support	0-0.08	4 Reject, 0 Support

3.4.2 Gene sampling

Results of Bayes Factor-based tests of monophyly were generally consistent across genes, except for the MC1R gene. This gene also exhibited an anomalous pattern when phase resolution was estimated; a pattern inconsistent coalescent-based anomalous lineage sorting, given the depth of phylogeny being investigated. Therefore, as a qualitative measure of its contribution to the multi-locus analyses, we re-estimated the concatenated phylogeny and species tree, excluding the MC1R data. Interestingly, the topologies changed and overall support (average BPP across all nodes) improved in both analyses, and *Virginia* changed from a well-supported paraphyletic pair to a well-supported polyphyletic pair (Fig. 3.2). The MC1R gene is part of the pigmentation pathway, a phenotypic characteristic that is often adaptive and broadly under strong selection; MC1R has been suggested to be adaptive and under selection in reptiles (Rosenblum et al., 2004). Additionally, patterns of pigmentation are often convergent among snakes and members of Thamnophiini are no exception. Though relatively easy to amplify and sequence, we would recommend that this gene be used with caution in phylogenetic and phylogeographic studies without incorporation of a more robust understanding of its evolution.

3.4.3 Evolutionary and Taxonomic Implications

Rossman (1963) described the crayfish snakes as sharing many morphological characteristics but displaying two distinct types: the pair with more standard dentition, *Regina grahamii* and *R. septemvittata*, which feed on recently molted crayfish, and the more extremely derived type (*R. alleni* and *R. rigida*) with chisel-like, kinetic teeth and specialized feeding behavior (Franz, 1977; Godley, 1985; Myers, 1987). Our data lend no support to the former type as a valid taxonomic group; however there was no outright rejection based on posterior probability. With morphological and allozyme-based evidence (Lawson, 1985) supporting this group, we are hesitant to suggest that they have independently evolved along ecological and morphological pathways without further study. Further, if the relationship in the concatenated estimate including MC1R is accurate (Fig. 1a), their similarities may represent shared ancestral characters. The latter group is supported, but with the inclusion of the black swamp snake (*Seminatrix pygaea*) as sister to *R. rigida*. Interestingly, this indicates a shift away from a specialized feeding ecology, and accompanying morphology, to a generalized diet in the swamp snake, which include amphibians, fish, and a variety of invertebrates (ref).

The earth snakes are represented by two small, gray, fossorial species, with largely overlapping ranges and subsisting on earthworms (Conant and Collins, 1991). Neither our nor the previous allozyme study support monophyly of this group, though, similar to the abovementioned case, the concatenated analysis including MC1R (Fig. 1a) suggests that they may share ancestral characters as basal members of the clade containing *Clonophis* and "*Liodytes*." These findings highlight convergent evolution in feeding morphology similar to that observed in other natricine snakes (e.g,. Vincent et al., 2009).

Our data lend support to the previous argument that crayfish predation arose more than once among Thamnophiini (Table 2). Advances genomic data collection and analytical methodology will facilitate investigation into the timing of divergence and relationships among these taxa, allowing for more robust models of trait origins. Our data support the findings of Alfaro and Arnold (2001), and we support the resurrection of the genus *Liodytes* for the currently recognized *Regina alleni* and *rigida*, with *Seminatrix* nested within this genus. In the case of the earth snakes, there was virtually no support for but ample rejection of their monophyly. Based this evidence, we suggest the resurrection of the genus *Haldea* (Baird and Girard, 1853) for the currently recognized *Virginia striatula*.

CHAPTER 4.
MULTILOCUS PHYLOGENY OF THAMNOPHIINI AND THE LABILITY OF PREY CHOICE

4.1. INTRODUCTION

While molecular phylogenetic estimates provide vital data pertaining to the relationships among organisms, the utility of these estimates is enhanced when they serve as the basis for downstream analyses. For example, comparative methods (i.e., the optimization of organismal features on the phylogenetic estimate) can lead to insight regarding phenotypic evolution, particularly when phylogenetic independent contrasts (Felsenstein, 1985) are utilized. Understanding the timing of evolution by tracking rates of cladogensis can improve our understanding of species diversification. In addition to providing a historical context to interpret the evolution of organismal features, phylogenies aid the researcher in understanding branching patterns and identifying the factors that promoted diversification. When combined, analytical tools that track character state evolution and the diversification of lineages through time improve comprehension of both the pattern and process of evolutionary diversification. Here we apply these tools to the thamnophiine snakes, an understudied group of vertebrates that have diversified into a variety of feeding niches.

Snakes are a diverse clade of vertebrates (>2900 extant species) that occupy a wide range of habitats and ecological niches despite obligate carnivory. Major functional adaptations in snakes can be categorized into two (none exclusive) types: locomotion and feeding. Arguably, the most important adaptation during the snake radiation was the evolution of the relaxed jaw articulations allowing for consumption of larger prey items; this synapomorphy is shared by all macrostomatans, which account for greater than 85% of extant snake diversity. This innovation has enabled further adaptations in the feeding apparatus, including the highly derived venom delivery systems of Elapidae and Viperidae.

Within the macrostomatans, the Thamnophiini (58 currently-recognized species) represent the natricine subfamily of colubrid snakes in the western hemisphere. This large radiation is traditionally classified (based largely on morphology) into nine genera that span from Canada to Costa Rica and occupy a variety of montane to estuarine habitats. Many thamnophiine snakes are diet specialists, including those whose prey choice is restricted to soft prey, such as earthworms and slugs; and those that prefer hard prey, such as crayfish. Most species are closely associated with water, either as their primary habitat or as a source of prey for both aquatic and terrestrial foragers (Gibbons and Dorcas, 2004; Rossman et al., 1996). Molecular phylogenetic data has supported many previously hypothesized clades, but also suggested that several of the clades inferred from morphological data are paraphyletic. Notable examples of paraphyly include the inclusion of *Thamnophis validus* in the genus *Nerodia* and the non-monophyly of the genus *Regina*.

Previous genetic work that focused on the relationships among the thamnophiines include the allozyme studies by Lawson (1985) and de Queiroz and Lawson (1994) as well as the mitochondrial sequence based works of Alfaro and Arnold (2001) and de Queiroz et al (2002). Important generic-level taxonomic discoveries were made from each of these studies. For example, phylogenies presented by both Lawson (1985) and Alfaro and Arnold (2001) are inconsistent with the hypothesis that both *Regina* and *Virginia* are monophyletic, and evidence from de Queiroz et al. (2002) suggests that the distinctive mountain meadow snakes (genus *Adelophis*) are nested within the garter snakes (*Thamnophis*). Alfaro and Arnold (2001) designated three major lineages in the thamnophiine snakes: the garter snakes (*Thamnophis*), the water snakes (*Nerodia*, *Regina grahamii* and *R. septemvittata* and *Tropidoclonion*), and the semifossorial snakes (*Clonophis, Regina alleni* and *R. rigida, Seminatrix, Storeria*, and *Virginia*). Despite these findings, key taxanomic findings have not been confirmed using a multi-locus phylogeny of the North American natricine snakes. Since multiple loci are required to generate accurate estimates of phylogeny (Kim and Burghman, 1988; Townsend, 2007) , we seek to generate such an estimate and use it to evaluate the key sources of conflict between morphological data and previous molecular work. Finally, after we estimate relationships within the Thamnophiini, we examine the patterns of diversification through time and discuss adaptations in feeding habitat use in a systematic framework.

4.2. METHODS

4.2.1. Data collection

Tissues from 52 specimens representing 51 (50 ingroup + *Natrix natrix* as outgroup) species were obtained primarily from two museum collections (Appendix C Table 1). The outgroup is a representative of the European radiation of the subfamily Natricinae. DNA was extracted via a modified salt-saturation protocol (Aljanabi and Martinez, 1997), in which tissue was lysed using 300µL of PureGene Cell Lysis solution (QIAGEN catalog no. 158906) followed by overnight incubation with proteinase K (New England Biolabs no. P8102S).

From previous literature, we amplified five (one mitochondrial and four nuclear) coding loci for each individual (Table 4.1). Additionally, anonymous nuclear markers were developed for this study by screening a fragment library previously prepared for microsatellite discovery, following Glenn and Schable (2005). Initial fragments were selected that were determined not to contain variable number tandem repeat regions, as detected by the *abblast* function of repeatmasker (Smit et al., unpublished), We then developed primers using Primer3 (Rozen and Skaletsky, 2000) and screened for amplifiability in a four taxon test set (*Natrix*, *Nerodia*, *Storeria* and *Thamnophis*). Following amplification across all taxa, five fragments were ultimately selected for sequencing.

Polymerase chain reaction (PCR) amplification of fragments was performed via polymerase chain reaction with reagent proportions as follows: 0.4-1ng/µL tDNA, 0.4µM each primer, 0.2µM dNTPs, 1X Standard Taq reaction buffer (New England Biolabs) and

0.5 units of Taq DNA polymerase (New England Biolabs no. M0267) per 25µL final volume. Thermocycling conditions were optimized for primer melting temperature and target fragment length (Table 1). Bi-directional sequencing for both coding and anonymous fragments was performed using Big Dye v 3.1 (Applied Biosystems, Foster City, CA) following manufacturer's protocols. Sequencing was performed at LSU and Beckman Coulter (Danvers, MA), Sequences were then analyzed on an ABI 3130 genetic analyzer (Applied Biosystems) at the genomics center at LSU or at Beckman Coulter. Chromatograms were examined by eye and edited using Sequencher 4.8 (Gene Codes, Ann Arbor, MI). Alignment of loci was conducted using Muscle (Edgar, 2004a, b), under the default settings.

Table 4.1. Primer and thermocycler information for each locus. Shown are the Locus, the sequence of the primer (5'-3'), the Annealing temperature (T_A; Celsius), the time of elongation ($Time_E$), and the source of the primer.

Gene		Oligo (5'-3')	T_A	$Time_E$	Reference
BDNF	F	GACCATCCTTTTCCTKACTATGGTTATTTCATACTT	50	:30	Leache and McGuire (2006)
	R	CTATCTTCCCCTTTTAATGGTCAGTGTACAAAC			
FSHR	F	CCDGATGCCTTCAACCCVTGTGA	50	:30	Wiens et al. (2008)
	R	CCRAAYTTRCTYAGYARRATGA			
MC1R	F	TCAGCAACGTGGTGGA	50	:30	Austin et al. (2009)
	R	ATGAGGTAGAGGCTGAAGTA			
ND4	F	TGACTACCAAAAGCTCATGTAGAAGC	55	:30	Forstner et al. (1995)
	R	TTTTACTTGGATTTGCACCA			Skinner et al. (2006)
NT3	F	ATGTCCATCTTGTTTTATGTGATATTT	50	:30	Wiens et al. (2008)
	R	ACRAGTTTRTTGTTYTCTGAAGTC			
R35	F	TCTAAGTGTGGATGATYTGAT	50	:30	Fry et al. (2006)
	R	CATCATTGGRAGCCAAAGAA			
"E"	F	CTGGATCCATAGCTCCTGGT	52	:20	This study
	R	ATTTTCAACCCAGCTTTTGG			
"I"	F	GGGAAAAAGAGGGAAATTGG	52	:20	This study
	R	GTGAAGGGTTTTGGGTGTTG			
"K"	F	GCCACCCTGACACTAAAAACA	52	:20	This study
	R	TTCCTGGAAGATGGTTTTGC			
"M"	F	TGAATGAGGCTGCGAGATTA	52	:20	This study
	R	AGGGGAGCCAGGTGTAACTT			

4.2.2. Phylogeny and Divergence Dating

A Bayesian estimate of phylogeny was generated using BEAST 1.7.4 (Drummond et al., 2012). Prior to analysis, we selected site models for each locus using DT-ModSel (Minin et al., 2003) and PAUP* (Swofford, 2003). Optimal models for each locus were defined in BEAST, with the exception of models that contained both a gamma-distributed rates and invariant sites (G+I). Such models were simplified to use only the gamma distribution to describe rate variation due to the potential interference between variables (Yang, 2006). Each gene was allowed to evolve under an independent, uncorrelated relaxed lognormal clock, with each sample mean drawn from a uniform prior with range 0-100 to allow for differences in substitution rate among genes. Two identical MCMC runs of $2x10^8$ steps (sampling every $2x10^4$) were performed and posterior distributions of parameters were compared for convergence using Tracer 1.5 (Rambaut and Drummond, 2009). Additionally, we estimated a Bayesian tree separately for each locus using BEAST.

To estimate divergence times across lineages, we performed two additional MCMC runs in BEAST, allowing all loci to evolve under a single uncorrelated relaxed lognormal clock, and calibrating the analysis by incorporating two fossils representing the oldest known specimens of *Nerodia* and *Thamnophis*, both from the Medial Barstovian fossil age in the Miocene (Holman, 2000; Appendix C Table 2). Two identical chains were allowed to run until effective sample sizes were above 100, and we compared trace files of both runs to ensure that runs had converged.

We recognize the importance of testing differing models of evolution when conducting phylogenetic studies; in particular, a coalescent-based species tree approach may be equally or more appropriate for these data (Chapter 2). However, no single locus was represented by all taxa in our dataset due to difficulties with PCR amplification for certain loci in some taxa, and a drawback to coalescent-based methods is their inability to accept datasets when one or more species is missing all alleles at one or more loci. To explore whether the model of evolution would significantly affect the topological outcome, we sub-sampled the dataset to include only those individuals having alleles at all loci. We then conducted two independent runs for each model of evolution (concatenation, coalescent-based species tree using *BEAST) to assure convergence within each model. Finally, we compared topologies between models to assess discordance supported by high posterior probabilities. Topologies were discordant at two nodes, but neither node was strongly supported (BPP >0.95; Appendix C Figure 1).

4.2.3. Diversification Rates

Lineage-through-time plots and their associated analyses allow us to test hypotheses concerning patterns of diversification across a phylogeny. We used the LASER (Rabosky, 2006) and APE (Paradis et al., 2004) libraries implemented in R to assess patterns of diversification rates among the ingroup. Specifically, we can test whether the rate of diversification is higher or lower than expected at any depth across the tree by comparing it to a null distribution of expected rates based on the speciation rate of the input tree. A lineage through time plot of the ultrametric MCC tree (pruned to contain the ingroup using the *drop.tip* function) was compared to a null distribution of diversification simulated with *rbdtree.n*, under a Yule pure birth model with a constant diversification rate. To assess the fit of other speciation models to our data, we attempted to fit our data to a birth-death model of diversification using the *birthdeath* function in APE.

4.2.4. Ancestral Trait Estimation

In order to understand the evolution of feeding specialization, we employed two methods of ancestral state reconstruction implemented in Mesquite (Maddison and Maddison, 2011) and BayesTraits (Pagel et al., 2004). We focused on two multi-state ecological features among species: diet and habitat type (Table 4.2). In Mesquite, we used maximum likelihood-based stochastic evolution modeling, with rates estimated from the character matrices, to trace the diet and habitat ancestral states on the maximum clade credibility (MCC) tree obtained from the initial BEAST analysis. Then for each major lineage (*Nerodia*, *Thamnophis*, semi-fossorial clade) we conducted a BayesMultiState MCMC

($5x10^6$ generations, sampling every 100 steps) for each character using the *.trees* file obtained from BEAST; this model allowed us to integrate over uncertainty in the tree topology. Rates priors were obtained from a gamma-distributed reversible-jump hyperprior bound by zero and ten.

4.3. RESULTS

4.3.1. Sequence analysis

Proportion of variable sites differed across loci (Table 4.2), with and average length of 418 base pairs and an average of 72 variable sites per locus. The overall quality of the gene tree estimates (as ascertained by the nodal support values) varied among among gene trees (not shown) in a manner apparently correlated with the variability of each locus. Of note, inspection of the DNA sequence data and gene tree estimate of *Locus K* suggested that multiple loci were being amplified, as particular insertion/deletions were shared among polyphyletic assemblages spanning major lineages; therefore this locus was not included in the concatenated analysis. A fifth anonymous fragment which was sequenced (*Locus A*) was not included in the analysis, as it contained a single variable site in the ingroup which was heterozygous across all individuals, implying that a paralogous loci, each homozygous, were coamplified. Missing data was filled in where possible from Genbank data (Appendix C Table 1).

Table 4.2. Summary of DNA fragments analyzed. *bp*: final length of edited fragment used in study; *s*: total number variable sites.

Gene	bp	s	model
BDNF	557	31	K80 + G
FSHR	511	39	HKY + I
MC1R	435	56	HKY+G
ND4	614	270	K80+G
NT3	561	85	K80 + G
R35	645	81	HKY + G
E	196	44	HKY
I	209	34	K80
K	243	37	HKY + I
M	228	31	HKY + I
Total	4199	708	N/A

4.3.2 Phylogeny, Divergence and Diversification

The Bayesian estimate of phylogeny showed high support across most nodes in the tree (Figure 4.1); *Nerodia* was highly supported as monophyletic, with Bayesian posterior probabilities (BPP) of 1 *Thamnophis* was estimated as paraphyletic with *Adelophis* (BPP 1); a third clade, including the genera *Clonophis*, *Seminatrix*, *Storeria*, *Virginia*, and *Regina alleni* and *R. rigida* was highly supported (BPP 1). The remaining ingroup taxa, *Regina grahamii*, *R. septemvittata* and *Tropidoclonion lineatum*, fell outside the three

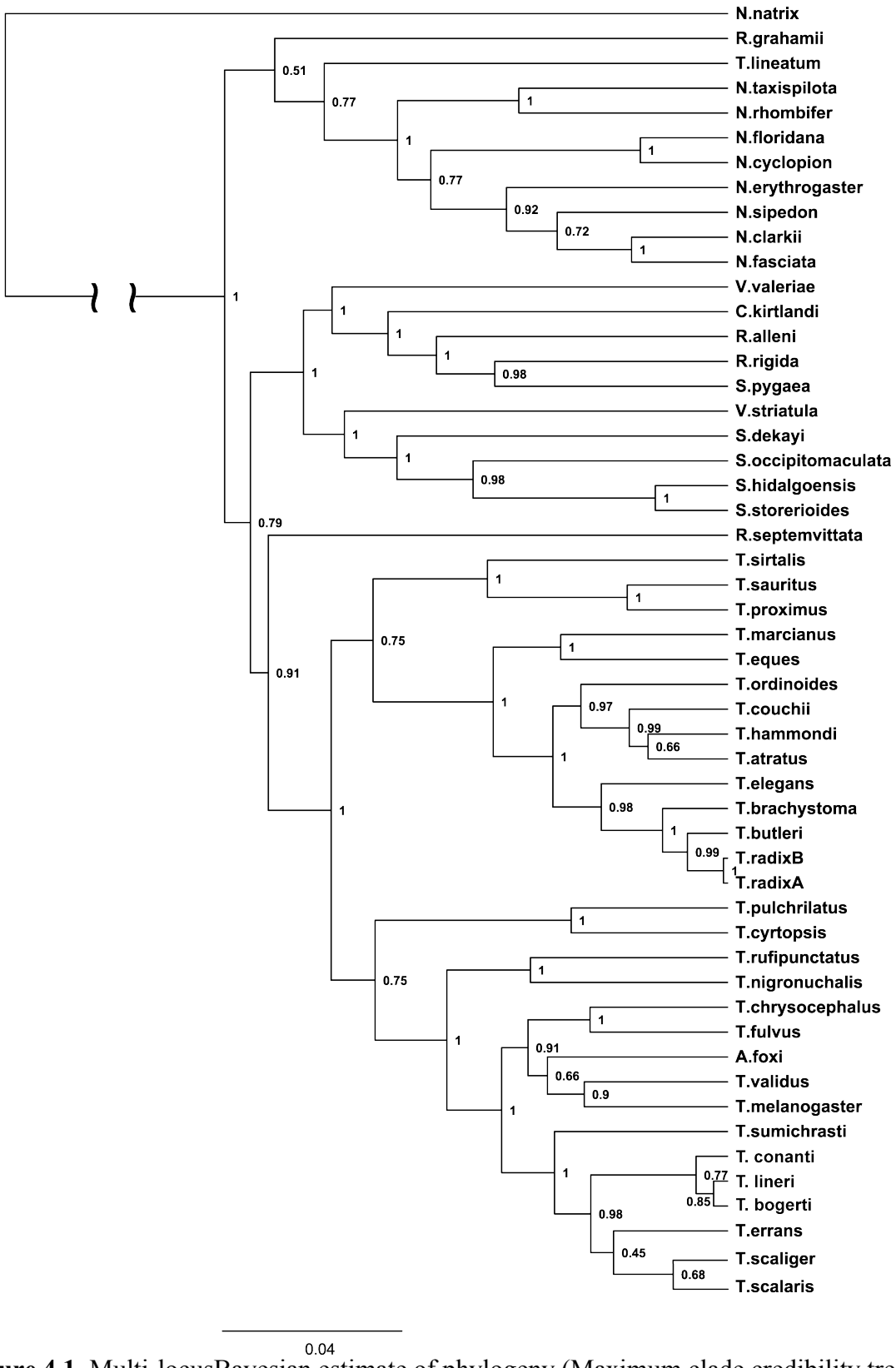


Figure 4.1. Multi-locusBayesian estimate of phylogeny (Maximum clade credibility tree) of Thamnophiini. Branch lengths based on ND4 rate estimated by BEAST.

aforementioned lineages, with only *R. septemvittata* marginally supported (BPP 0.92) in its placement, as sister to *Thamnophis*.

The estimated timing of diversification among the major lineages indicates that the ancestors of the thamnophiine snakes diversified quickly. The estimated divergence time of major lineages are broadly overlapping (Figure 4.2), with most of the diversification estimated to have occurred between during the Miocene (~14-11MYA). The ancestral node of Thamnophiini is estimated to have occurred 15 +/- 1MYA, later than the oldest known North American fossil attributed to the subfamily Natricinae (*Neonatrix elongata*). Consistent with the rapid estimated timing, the lineage through time plots (Figure 2) indicate that diversification increased between 6-2MYA, when there was a significant increase above the null rate of diversification ($p < 0.01$). For the lineage through time plot, a pure birth model provided a better fit to the data; the *birthdeath* function implemented in LASER estimated a death rate of 0. Of the 15 branching events estimated to have occurred during the Pliocene, 12 were within *Thamnophis.*

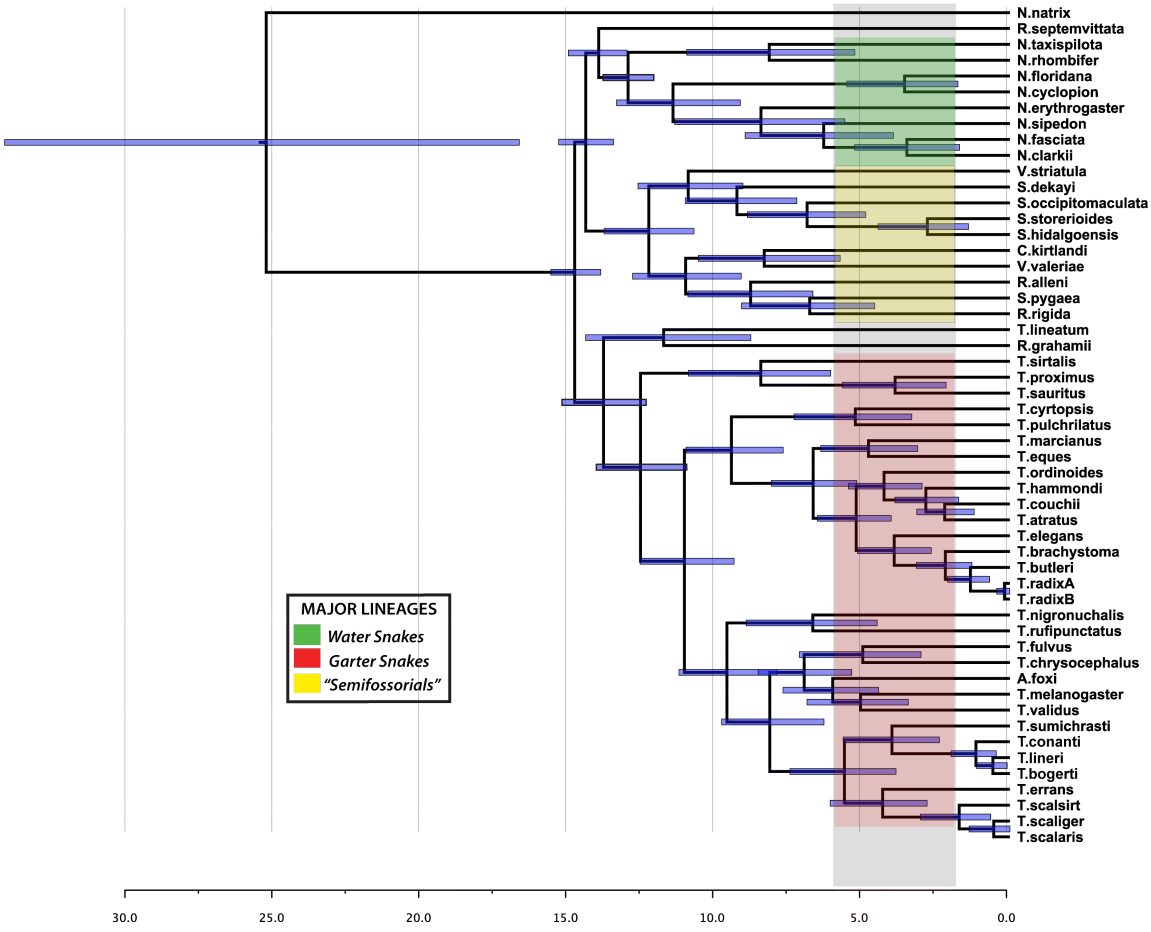


Figure 4.2. Chronogram of Thamnophiini, based on fossil calibrations of *Nerodia* and *Thamnophis*. Error bars denote node age 95% posterior distribution, and the major lineages (see text) are denoted with colored error bars (see legend). The estimated age of each node is given by the scale on the X-axis, in units of millions of years.

4.3.3 Ancestral state reconstruction

Reconstruction of ancestral diet was equivocal; Mesquite recovered no phylogenetic pattern of diet preference across the tree, and no internal nodes had significant preference for any of the five states (Figure 4.4). A stronger phylogenetic signal was evident in habitat preference, where the ancestor of Thamnophiini and *Nerodia* were reconstructed to prefer aquatic habitats. Ancestral states of *Thamnophis* and the semi-fossorial clade was ambiguous. Likewise, the BayesTraits-derived MCMC posteriors of each diet state showed no measurable deviation from equiprobability for each lineage. Aquatic habitat was estimated to be the ancestral state (0.958 of state posterior) in *Nerodia*. Ancestral habitat estimations of *Thamnophis* and the semi-fossorial clade were ambiguous, with "near-water" (0.468) and "semi-fossorial" (0.767) the most frequently sampled states, respectively.

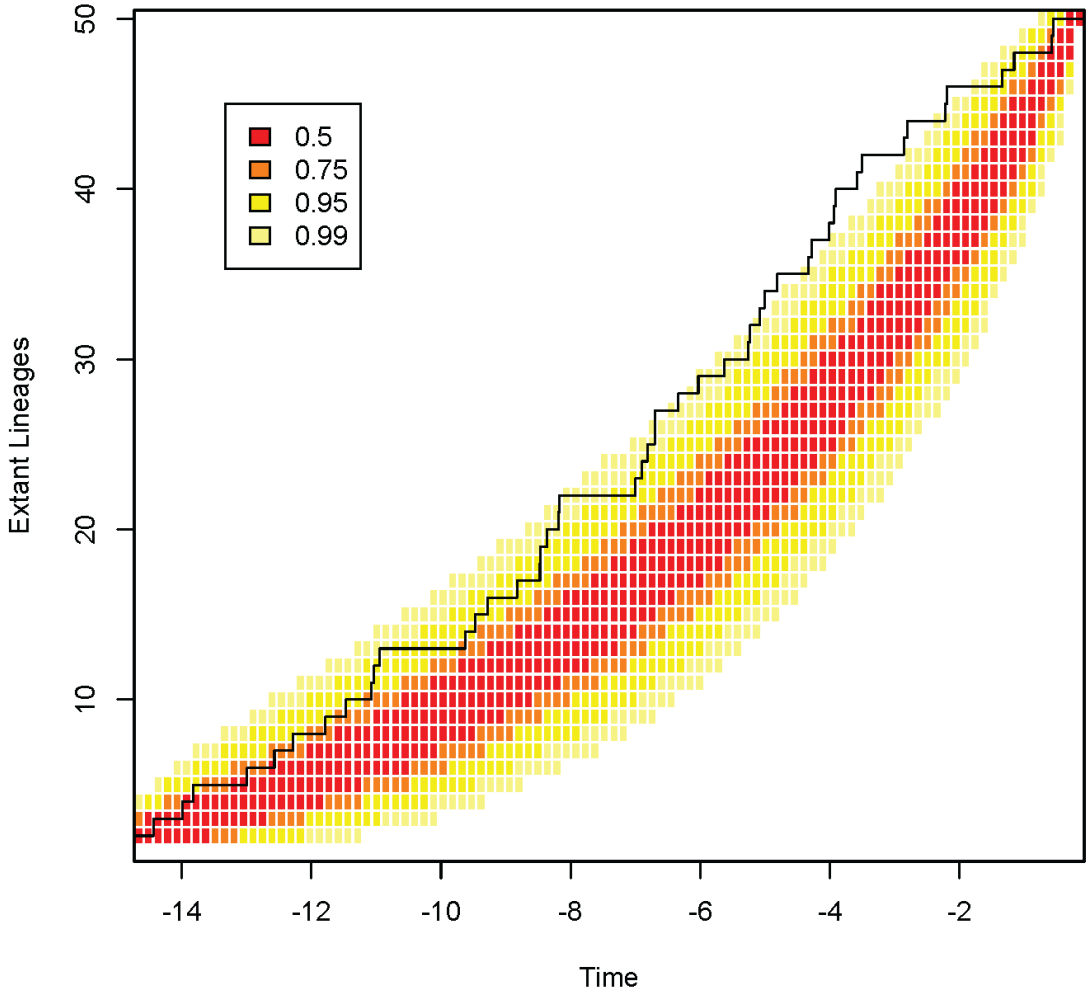


Figure 4.3. Lineage through time plot of for Thamnophiini. The Y-axis depicts the number of extant lineages, while the X-axis depicts time in units of millions of years. The null distribution is shown using the shaded heatmap, with a clear acceleration in the diversification of these lineages present in the recent.

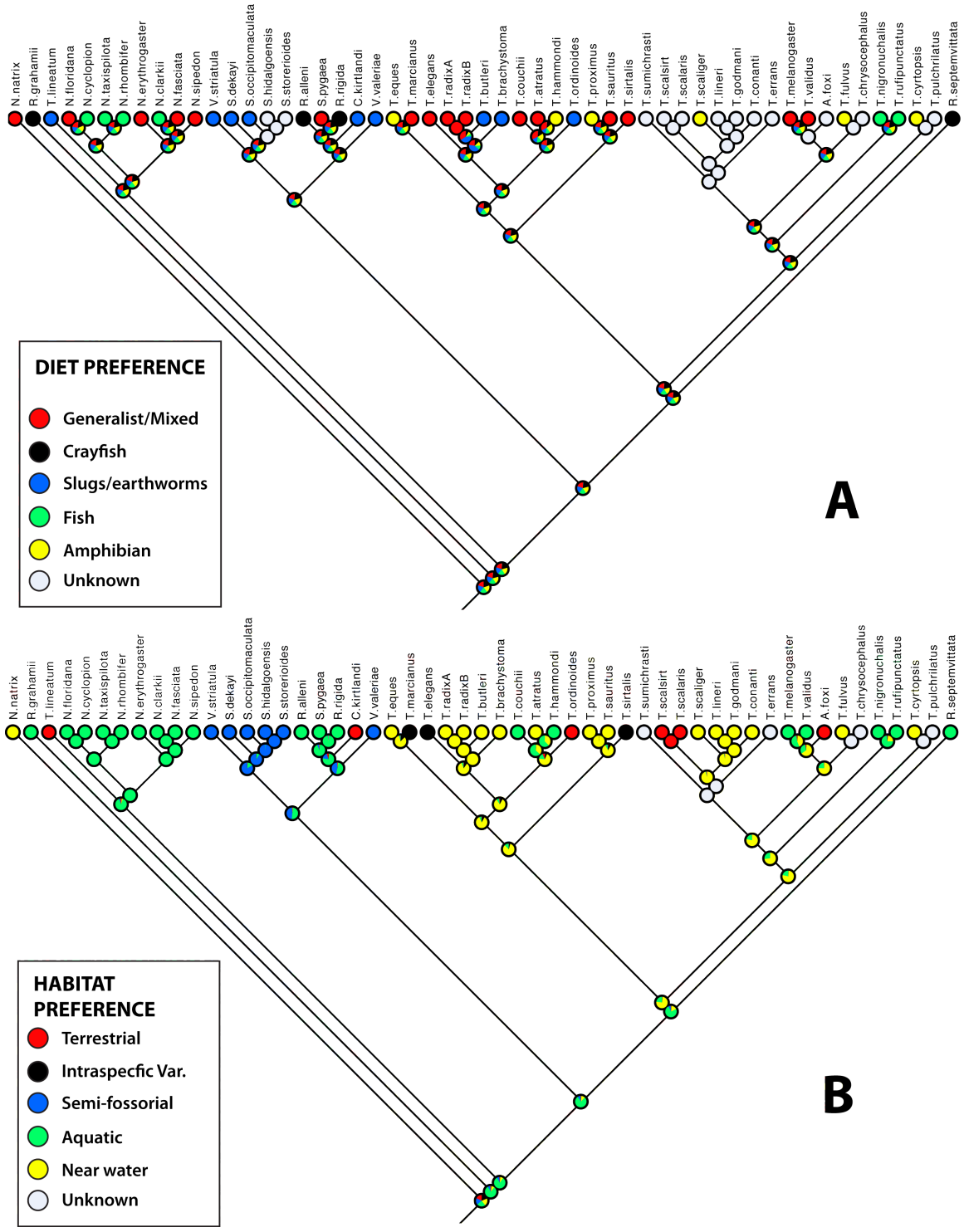


Figure 4.4. Ancestral state reconstruction of feeding and habitat type using maximum likelihood in Mesquite. A) Diet preferences; B) Habitat preferences. "Generalist/Mixed" combines species who feed on more that one prey type, or who exhibit population-level diet specialization. Data sources listed in Appendix C Table 3.

4.4. DISCUSSION

Prey choice is labile within thamnophiine snakes. Optimizations of diet characteristics demonstrate that feeding specializations such as stenophagy and specializations of tooth and feeding apparati have evolved multiple times, but we cannot reconstruct the ancestral diet because of the variety of feeding specializations that have evolved. Comparison with the clades sister to the Thamnophiini would be ideal to better optimize these characteristics, however less is known about many of the other natricine taxa when compared to Thamnophiini, both in phylogeny and diet preference. This inference is supported by phenotypic characteristics tied to diet, such as feeding strategy and tooth surface morphology, which also occur throughout the phylogeny. For example, Herrel et al. (2008) and Vincent et al. (2009) showed that head shape and feeding strategy ("sweeping" vs "striking") was labile within and across genera (*Thamnophis* and *Nerodia*), with strategies showing convergence among non-sister taxa that share prey choice. In addition, Britt et al. (2009) suggested that tooth surface morphology was also to be labile and convergent. One of the most extreme examples of adaptation of tooth surface morphology within the group is that shared by the crayfishs nakes *Regina* ("*Liodytes*") *alleni* and *R. rigida*. Virtually all genetic evidence collected to date suggests that this group is paraphyletic, and that the black swamp snake (*Seminatrix pygaea*) is sister to *R. rigida*. The swamp snake has adopted a more general prey choice, and microscopic examination of skeletal specimens by JDM could find no evidence of the medial compression or chisel-like appearance exhibited by its closest extant relatives. Prey choice can even vary intraspecifically among thamnophiines, such that individual populations become prey specialists, such as in the cases of *Thamnophis elegans* (e.g., Arnold, 1977), *T. melanogaster* (Manjarrez, 2005), and *T. sirtalis* (Britt et al., 2006). Unlike prey choice, habitat preference is more easily estimated for most ancestral nodes using either ML and Bayesian methods, with most ancestors estimated to prefer an aquatic or "near water" habitat. This is consistent with habitat choice found in Narticine radiations across the world.

4.4.2 Biogeography and lineage ages

Within *Thamnophis*, the twelve branching events estimated to have occurred during a period of increased diversification in the Pliocene occurred with two separate major lineages within the genus, whose ancestral node age range does not overlap with the branching events in the lineages, suggesting two independent radiations. Moreover, the bulk of the current distributions of these lineages are centered on different regions of the continent. Despite an increase in diversification over this time period, these lines of evidence are counter to a hypothesis of single rapid adaptive radiation. Burbrink and Pyron (2010) also found an increase above expected diversification, then a decrease in rate of speciation during the Pliocene in the snake tribe Lampropeltini; both our and their data are inconsistent with the Pleistocene speciation phenomenon seen in other North American taxa (Bermingham et al., 1992; Knowles, 2000; Levsen et al., 2012).

4.4.3 Phylogeny and Diversification

A primary goal of this study was to develop the most comprehensive phylogenetic estimate of the North American Natricine snakes, we met this goal by including novel genetic data from eight independently evolving genetic loci. The continual addition of taxa and characters to estimates of phylogeny serve to improve our model of the relationships of these organisms, and more broadly our understanding of the nature of cladogenesis. Of equal importance is that a more robustly-estimated phylogeny often contains less uncertainty and can bolster statistical confidence of studies, such as diversification and ancestral state estimation, that incorporate these estimate. In our case, inclusion of multiple nuclear loci allow us to gather information from regions of the genome that are evolving at different rates, increasing the phylogenetic informativeness of our dataset for nodes across the depth of the tree (Townsend, 2007). Our results largely agree with the findings of the previously published molecular-based studies, with exceptions noted below. Note that we have partitioned our discussion of the taxonomic implications of this work following the designations of Alfaro and Arnold (2001).

Garter Snakes.—Three well-supported clades were recovered within *Thamnophis*: two broadly-distributed clades, and one composed of species found only in México, Guatemala and Honduras. Our samples also include three of the most recent additions to *Thamnophis*: *T. lineri* and *T. conanti*, populations of *T. godmani* which were elevated to species status based on allopatry and morphological evidence (Rossman and Burbrink, 2005). Genetic divergence was evident, but lower than expected for distinct species (<1% uncorrected pairwise sequence divergence for each comparison). However, these species are each represented a single individual in our study; more thorough genetic sampling would be needed to approriately characterize the status of these lineages. Fox's Mountain Meadow Snake (*Adelophis foxi*; Rossman and Blaney, 1968), was strongly supported (BPP >0.95 at four internodes) as nested within *Thamnophis*. Despite these results we are hesitant to make the suggestion that the genus *Adelophis* be synonymized with *Thamnophis*, because this taxon is represented in our study by the specimen (LSUMZ 40848) used in the de Querioz et al. (2002) and thus subject to the same caveats as discussed by de Queiroz et al. However, our DNA was processed from a different aliquot of tissue, dispelling the possibility of bias due to PCR contamination (as discussed in de Querioz et al.).

Water Snakes.— As mentioned above, *Nerodia* was estimated as monophyletic with strong support. Our findings disagree with the relationships among species within *Nerodia* agree with those estimated by Alfaro and Arnold (2001); specifically, supported nodes from our estimate are in conflict with the hypothesis that *Regina grahamii*, *R. septemvittata* and *Tropidoclion* are nested within *Nerodia*. This finding is not particularly surprising, given the lack of strong support for this particular grouping in the former study. However, we cannot reject the hypothesis that *Nerodia* and these three taxa form a monophyletic assemblage, though McVay and Carstens (Chapter 3) did reject a sister relationship of *R. grahamii* and *R. septemvittata*, based on gene-by-gene tests of monophyly. This suggests that there are four independent origins of crayfish predation: the *Liodytes* clade, *R. grahamii*, *R. septemvittata*, and a population of *Thamnophis*

melanogaster in México (Manjarrez, 2005). Missing from our study is *N. harteri*, recovered as sister to *N. sipedon* by Alfaro and Arnold.

Semi-fossorial clade.—Though including more species than Alfaro and Arnold, we estimated a phylogeny consistent with their findings, including the paraphyletic nature of "*Liodytes*." Our findings are also consistent with McVay and Carstens (Chapter 3), who rejected the monophyly of *Virginia*, based on multiple gene tree-based tests of monophyly. We recovered with high support that *V. valeriae* has a sister relationship to *Clonophis*, "*Liodytes*" and *Seminatrix*, and *V. striatula* as sister to *Storeria*. Our study includes, for the first time, all currently-recognized species of *Storeria*. Interestingly, *S. hidalgoensis* is supported by our data as sister to the other species endemic to México, *S. storerioides*; the former was initially considered to be of *S. occipitomaculata* (Massachusetts. Zoological and Botanical Survey. et al., 1839), and its taxonomic status has been debated (Flores-Villela, 1993; Taylor, 1942; Trapido, 1944).

4.4.3 Future Research

Our phylogenetic estimate is dependent on the data, and while we have collected the largest dataset to date in Thamnophinii, we anticipate that the data available for phylogeny reconstruction in this group will increase dramatically as researchers incorporate high throughput sequencing (e.g., McCormack et al. 2013). While genome-scale sequencing will likely improve our understanding of the broader relationships, these advances can also contribute to the need for finer scale genetic exploration, both interspecific and among populations. To date, phylogeographic and/or population genetic results have been published for only a handful of the currently recognized species in this group, including *T. nigronuchalis* and *T. rufipunctatus* (Wood et al., 2011), *T. sirtalis*, (Janzen et al., 2002), *T. elegans* , *T. proximus* (Allen, 2005), *T. validus* (de Queiroz and Lawson, 2008), *N. clarkii* (Jansen et al., 2008) *N. erythrogaster* (Makowsky et al., 2010), *N. rhombifer* (Brandley et al., 2010). With increased understanding of gemone structure, use of high throughput sequencing can simultaneously address population genetic and functionally evolutionary questions. For example, we used a double digest RADseq approach to assess the history of hybrid zones between *N. clarkii* and *N. fasciata*, and between *N. fasciata* and *N. sipedon* in Louisiana (Appendix 4.4); mapping these data to a genome in the future may allow us better understand the ecological nature of their divergence prior to secondary contact. Virtually nothing is known about the phylogeography of any members of Thamnophiini outside of *Nerodia* and *Thamnophis*. Of equal importance is the need for continued research into the ecology, morphology and behavior of this group. These data will be critical in developing a complete understanding of the Thamnophiini because they will (i) help to validate species boundaries and (ii) hybrid zones, while leading to an increased understanding of how habitat and climatic change have influenced the evolution of this group.

CHAPTER 5.
CONCLUSIONS

5.1. Summary

My fascination with living world drove me to pursue a career in biology. With an inherent desire to organize and catalog items of all varieties, I gravitated towards systematic biology, and particularly the study of herptiles. While the Thamnophiini, including the North American water snakes and garter snakes, are a relatively common group that has been well-studied, its phylogeny was not fully known at the inception of my dissertation. Investigations into the Thamnophiini have been facilitated through the analysis of DNA sequence. The predictability of this molecule allows us to model its evolution and phylogeny in a manner less prone to the potential biases of analyzing morphology and other types of biological data. The results of my research improve our understanding of the evolution and phylogeny of Thamnophiini. The focus of this work: the study of thamnophiine evolution in a robust methodological framework, provides the foundation of my future research endeavors. I also hope that the methods and results outlined here will facilitate the future research of others, within and beyond the thamnophiine snakes.

Chapter 2.—I demonstrate a useful and direct approach to choosing among the two dominant phylogenetic models; concatenation and species tree estimation. Central to our description of the issues related to choosing among these models is the assumption that it is important to have an *a priori* expectation of model performance in order to avoid a *post hoc* evaluation of the phylogenies. We also contend that the optimal sampling design differs for these competing models; for our data we show clearly that coalescent processes are likely to produce incongruence across loci and therefore future efforts will be focused on increasing the number of individuals included in the analysis.

Chapter 3.—I focus on the evolution and taxonomy of thamnophiine snakes. Similar to those data collected for the previous chapter, I encountered uncertainty in gene tree estimates due to low numbers of variable sites among loci. Despite this, I was able to make statistically supported taxonomic conclusions, owing in part to recently described approaches to making valid comparisons among Bayesian marginal likelihood estimates (Xie et al., 2011). I conclude from multiple types of comparisons that the genera *Virginia* and *Regina* are not monophyletic (Table 5.1), and that the taxonomy should be updated to reflect this.

Chapter 4.— I develop the most comprehensive phylogenetic estimate of the North American Natricine snakes to date (Table 5.2). Inclusion of multiple multiple nuclear loci allowed me to gather information from regions of the genome that are evolving at different rates, increasing the phylogenetic informativeness of our dataset for nodes across the depth of the tree (Townsend, 2007). The incorporation of fossil data and analysis of diversification rates across the tree allowed me to make conclusions about the evolutionary history of this group: the origin of all three major lineages likely occurred during the Miocene (~14-11MYA), and increased diversification occurred most notable

in the garter snakes during the Pliocene (~6-2MYA). Also, I show that both habitat and diet preference are labile within the thamnophiine snakes, and that specializations, such as crayfish predation have independently arisen multiple times across the phylogeny. This lability, along with the convergent and labile nature of superficial morphology (similarities in coloration patter, bauplan) has contributed to the historic confusion surrounding the taxonomy of this group.

Table 5.1. Summary of taxonomic evidence for contentious monophyletic groups in Thamnophiini. The red "x" and green check indicate statistically supported rejection or support for each group; "-" indicates no statistical support.

		LINES OF EVIDENCE			
Taxonomic Grouping	Species Tree	Concatenation	Gene Trees	Posteriors	BF Constraint Tests
"Earth Snakes"	-	×	2 ×, 0 ✓	0-0.03	4 ×, 0 ✓
"Crayfish Snakes"	×	×	3 ×, 0 ✓	0 in all genes	5 ×
Liodytes	×	×	1 ×, 0 ✓	0-0.11	2 ×, 2 ✓
Liodytes + *Seminatrix*	-	✓	1 ×, 0 ✓	0.007-0.13	1 ×, 4 ✓
R. grahamii + *R. septemvittata*	-	-	0 ×, 0 ✓	0-0.08	4 ×, 0 ✓

Table 5.2. Previous and current phylogenies of thamnophiine snakes based on DNA sequence data. *indicates number of independently sorting loci.

Study	Year	Data type	Number of ingroup species	Loci*
Alfaro and Arnold	2001	Mitochondrial Sequence	29	1
de Queiroz et al.	2002	Mitochondrial Sequence	33	1
This study	2013	Mitochondiral + Nuclear	50	8

5.2. Future directions of research

High throughput sequencing.—Recent advances in DNA sequencing technology have allowed researchers to sequence DNA in a massively parallel manner. These technologies, combined with techniques that target homologous regions of the genome across individuals (as the sequencers are designed for capture effectively random portions of the genome), are beginning to allow researchers to address questions I have posed using thousands instead of tens of independently sorting loci (Davey et al., 2011; McCormack and Faircloth, 2013; McCormack et al., 2013; Peterson et al., 2012). Data of this magnitude will doubtlessly facilitate our ability to understand the evolution of thamnophiine snakes (and other groups), from both a systematic and genomic perspective.

Lingering taxonomic questions.—In addition to the contentious taxonomic designations addressed in my dissertation, there a number of populations whose taxonomic statuses remain in question. Mitochondrial sequence data (McVay, unpublished) suggests deep splits among populations of both *Virginia* ("*Haldea*") *striatula* and *V. valeriae*. Additionally, a montane population of *V. valeriae* ("*V. pulchra*") was considered distinct by Richmond (1954) but has yet to be characterized genetically. Another example of a genetically uncharacterized putative species is the distinct Florida peninsular population of the Brown Snake, "*Storeria* (*dekayi*) *victa*" (Hay, 1892). Both cases warrant

examination; the former is of some urgency, given the conservation status of *V. pulchra* (small range and declining habitat; listed as endangered in Maryland). More generally, studies are needed to address the status (and definition) of subspecies across all genera in this group.

Equally important is the need for natural history and ecological studies in this group, particularly for those species underrepresented in the literature. The need also exists to examine in more detail traits associated with diet, including morphology, kinematics and behavior (Alfaro, 2003; Flores-Villela, 1993; Herrel et al., 2008; Taylor, 1942; Trapido, 1944; Vincent et al., 2009; Vincent et al., 2006), in a broader phylogenetic context; I believe that the results of my dissertation will facilitate these types of studies.

5.3. Conclusions

Through my dissertation research, I have improved our understanding of the phylogeny of thamnophiine snakes by incorporating sequence data from multiple nuclear loci, and including samples of taxa previously unrepresented in molecular studies. My results provided more evidence for rejection of the crayfish snakes, *Regina*, as a monophyletic assemblage, and from the inclusion of both recognized species of *Virginia*, I was able to conclude that this genus is polyphyletic as well. For the first time, I have included all know species of the genus *Storeria* in a molecular systematic study, and have confirmed and have rejected the hypothesis that *S. hidalgoensis* is a subspecies of *S. occipitomaculata* (Trapido, 1944), as it is supported to be sister to *S. storerioides*. I have confirmed that feeding is a labile trait within this group, with stenophagy having evolved multiple times; this is consistent with the broader pattern of convergence in ecologies among the major global radiations of natricine snakes. I have also estimated the first fossil-calibrated chronogram for this group, estimating a Miocene origin and Pliocene expansion of diversity. Further work is required to place *Regina grahamii*, *R. septemvittata* and *Tropidoclonion* with confidence within the phylogeny. Additionally, the placement of the rare meadow snake (*Adelophis*) requires confirmation with additional specimens and inclusion of the species not in this study, *A. copei*.

REFERENCES

Alfaro, M.E., 2003. Sweeping and striking: a kinematic study of the trunk during prey capture in three thamnophiine snakes. J Exp Biol 206, 2381-2392.

Alfaro, M.E., Arnold, S.J., 2001. Molecular systematics and evolution of Regina and the Thamnophiine snakes. Mol Phylogenet Evol 21, 408-423.

Aljanabi, S.M., Martinez, I., 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. Nucleic Acids Res 25, 4692-4693.

Allen, L., 2005. Phylogeography of five susbspecies of western ribbon snakes (*Thamnophis proximus*) in the U.S. and middle America. The University of Texas at Arlington, Arlington.

Baird, S.F., Girard, C., 1853. Catalogue of North American reptiles in the Smithsonian Institution. Part 1. Serpents. Smithsonian Institution, Washington.

Bermingham, E., Rohwer, S., Freeman, S., Wood, C., 1992. Vicariance biogeography in the Pleistocene and speciation in North American wood warblers: a test of Mengel's model. Proc Natl Acad Sci U S A 89, 6624-6628.

Brandley, M.C., Guiher, T.J., Pyron, R.A., Winne, C.T., Burbrink, F.T., 2010. Does dispersal across an aquatic geographic barrier obscure phylogeographic structure in the diamond-backed watersnake (*Nerodia rhombifer*)? Mol Phylogenet Evol 57, 552-560.

Brandley, M.C., Huelsenbeck, J.P., Wiens, J.J., 2008. Rates and patterns in the evolution of snake-like body form in squamate reptiles: evidence for repeated re-evolution of lost digits and long-term persistence of intermediate body forms. Evolution 62, 2042-2064.

Britt, E.J., Hicks, J.W., Bennett, A.F., 2006. The energetic consequences of dietary specialization in populations of the garter snake, *Thamnophis elegans*. J Exp Biol 209, 3164-3169.

Burbrink, F.T., Pyron, R.A., 2010. How does ecological opportunity influence rates of speciation, extinction, and morphological diversification in New World ratsnakes (tribe Lampropeltini)? Evolution 64, 934-943.

Collins, J.T., Taggart, T.W., 2002. Standard common and current scientific names for North American amphibians, turtles, reptiles & crocodilians. Center for North American Herpetology Lawrence, Kansas.

Conant, R., Collins, J.T. 1991. A field guide to reptiles and amphibians: eastern and central North America. The Peterson field guide series 12. Houghton Mifflin, Boston.

Cope, E.D., 1885. Twelfth contribution to the herpetology of tropical America. P Am Philos Soc 22, 167-194.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12, 499-510.

de Queiroz, A., Lawson, R., 1994. Phylogenetic relationships of the garter snakes based on DNA sequence and allozyme variation. Biol J Linn Soc 53, 209-229.

de Queiroz, A., Lawson, R., 2008. A peninsula as an island: multiple forms of evidence for overwater colonization of Baja California by the gartersnake *Thamnophis validus*. Biol J Linn Soc 95, 409-424.

de Queiroz, A., Lawson, R., Lemos-Espinal, J.A., 2002. Phylogenetic relationships of North American garter snakes (*Thamnophis*) based on four mitochondrial genes: How much DNA sequence is enough? Mol Phylogenet Evol 22, 315-329.

Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. Plos Genetics 2, 762-768.

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24, 332-340.

Drummond, A.J., Ho, S.Y., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS biology 4, e88.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29, 1969-1973.

Edgar, R.C., 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. Bmc Bioinformatics 5, 113.

Edgar, R.C., 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792-1797.

Edwards, A.W., Cavalli-Sforza, L.L., 1964. Reconstruction of evolutionary trees. Phenetic and phylogenetic classification 6, 67-76.

Edwards, S.V., 2009. Is a New and General Theory of Molecular Systematics Emerging? Evolution 63, 1-19.

Felsenstein, J., 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet 25, 471-492.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17, 368-376.

Felsenstein, J., 1985. Phylogenies and the Comparative Method. Am Nat 125, 1-15.

Felsenstein, J., 2004. Inferring phytogenies. Sunderland, Massachusetts: Sinauer Associates 4.

Flores-Villela, O., 1993. Herpetofauna mexicana : lista anotada de las especies de anfibios y reptiles de México, cambios taxonómicos recientes, y nuevas especies = Annotated list of the species of amphibians and reptiles of Mexico, recent taxonomic changes, and new species. Carnegie Museum of Natural History, Pittsburgh.

Flot, J.F., 2010. SEQPHASE: a web tool for interconverting phase input/output files and fasta sequence alignments. Mol Ecol Res 10, 162-166.

Franz, R. 1977. Observations on the food, feeding behavior, and parasites of the striped swamp snake, *Regina alleni*. Herpetologica 33, 91-94.

Garman, S., 1883. The reptiles and batrachians of North America. Memoirs of the Museum of Comparative Zoology 8, 1-185.

Gibbons, J.W., Dorcas, M.E., 2004. North American watersnakes : a natural history. University of Oklahoma, Norman.

Glenn, T.C., Schable, N.A., 2005. Isolating microsatellite DNA loci. Methods Enzymol 395, 202-222.

Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22, 160-174.

Hay, O.P., 1892. Description of a supposed new species of Storeria from Florida, *Storeria Victa*. Science (New York, NY) 19, 199.

Hennig, W., 1950. Grundzüge einer Theorie der phylogenetischen Systematik. Deutscher Zentralverlag.

Herrel, A., Vincent, S.E., Alfaro, M.E., S, V.A.N.W., Vanhooydonck, B., Irschick, D.J., 2008. Morphological convergence as a consequence of extreme functional demands: examples from the feeding system of natricine snakes. J Evol Biol 21, 1438-1448.

Holman, J.A., 2000. Fossil snakes of North America : origin, evolution, distribution, paleoecology. Indiana University Press, Bloomington.

Hull, D.L., 1990. Science as a process: an evolutionary account of the social and conceptual development of science. University of Chicago Press.

Jansen, K.P., Mushinsky, H.R., Karl, S.A., 2008. Population genetics of the mangrove salt marsh snake, *Nerodia clarkii* compressicauda, in a linear, fragmented habitat Cons Genet 9, 410-410.

Janzen, F.J., Krenz, J.G., Haselkorn, T.S., Brodie, E.D., 2002. Molecular phylogeography of common garter snakes (*Thamnophis sirtalis*) in western North America: implications for regional historical forces. Mol Ecol 11, 1739-1751.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian protein metabolism. Academic press, pp. 21-132.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J Am Stat Assoc 90, 773-795.

Kim, J., Burgman, M.A., 1988. Accuracy of phylogenetic-estimation methods under unequal evolutionary rates. Evolution, 596-602.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16, 111-120.

Kingman, J.F.C., 1982. The coalescent. Stoch Proc Appl 13, 235-248.

Knowles, L.L., 2000. Tests of pleistocene speciation in montane grasshoppers (genus *Melanoplus*) from the sky islands of western North America. Evolution 54, 1337-1348.

Kubatko, L.S., Carstens, B.C., Knowles, L.L., 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25, 971-973.

Lawson, R., 1985. Molecular studies of thamnophiine snakes. Louisiana State University.

Levsen, N.D., Tiffin, P., Olson, M.S., 2012. Pleistocene speciation in the genus *Populus* (salicaceae). Syst Biol 61, 401-412.

Li, W., 1997. Molecular evolution. Sinauer Associates Incorporated.

Maddison, W., 1997. Gene trees in species trees. Syst Biol 46, 523-536.

Maddison, W.P., Maddison, D.R., 2011. Mesquite: a modular system for evolutionary analysis.

Makowsky, R., Marshall, J.C., Jr., McVay, J., Chippindale, P.T., Rissler, L.J., 2010. Phylogeographic analysis and environmental niche modeling of the plain-bellied watersnake (*Nerodia erythrogaster*) reveals low levels of genetic and ecological differentiation. Mol Phylogenet Evol 55, 985-995.

Manjarrez, J., 2005. Posible invasión de unnicho alimentario nuevo y microevolución en una especie mexicana de serpiente. Ciencia Ergo Sum 12, 275-281.

Massachusetts. Zoological and Botanical Survey., Storer, D.H., Peabody, W.B.O., 1839. Reports on the fishes, reptiles and birds of Massachusetts. Dutton and Wentworth, State Printers, Boston.

McCormack, J.E., Faircloth, B.C., 2013. Next-generation phylogenetics takes root. Mol Ecol 22, 19-21.

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., Brumfield, R.T., 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol 66, 526-538.

Minin, V., Abdo, Z., Joyce, P., Sullivan, J., 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst Biol 52, 674-683.

Notredame, C., 2007. Recent evolutions of multiple sequence alignment algorithms. Plos Comput Biol 3, e123.

Pagel, M., Meade, A., Barker, D., 2004. Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53, 673-684.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289-290.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7, e37135.

Posada, D., 2008. jModelTest: phylogenetic model averaging. Mol Biol Evol 25, 1253-1256.

Posada, D., Crandall, K., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817 - 818.

Price, R.M., 1983. Microdermatoglyphics: the *Liodytes-Regina* problem. . J Herpetol 17, 292-294.

Pyron, R.A., Burbrink, F.T., 2009. Body size as a primary determinant of ecomorphological diversification and the evolution of mimicry in the lampropeltinine snakes (Serpentes: Colubridae). J Evol Biol 22, 2057-2067.

Rabosky, D.L., 2006. LASER: a maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. Evol Bioinform Online 2, 273-276.

Rambaut, A., Drummond, A.J., 2009. Tracer v 1.5.

Richmond, N.D., 1954. The ground snake Haldea valeriae in Pennsylvania and West Virginia with description of a new subspecies. Annals of the Carnegie Museum 33, 251-260.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61, 539-542.

Rosenblum, E.B., Hoekstra, H.E., Nachman, M.W., 2004. Adaptive reptile color variation and the evolution of the Mc1r gene. Evolution 58, 1794-1808.

Rossman, D.A., 1963. Relationships and taxonomic status of the North American natricine snake genera *Liodytes*, *Regina* and *Clonophis*. Louisiana State University, Baton Rouge.

Rossman, D.A., 1985. *Liodytes* resurrected, reexamined, and reinterred. J Herpetol 19, 169-171.

Rossman, D.A., Blaney, R.M., 1968. A new Natricine snake of the genus *Adelophis* from western Mexico. Occasional Papers of the Museum of Zoology Louisiana State University 35, 1-12.

Rossman, D.A., Burbrink, F.T., 2005. Species limits withing the Mexican garter snakes of the *Thamnophis godmani* complex. Occasional Papers of the Museum of Natural Science Louisiana State University 79, 1-44.

Rossman, D.A., Wallach, V. 1991. *Virginia* Baird and Girard. Earth snakes. Catalogue of American Amphibians and Reptiles No. 529, 1-4.

Rossman, D.A., Ford, N.B., Seigel, R.A., 1996. The garter snakes : evolution and ecology. University of Oklahoma Press, Norman.

Rozen, S., Skaletsky, H.J., 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz, S., Misener, S. (Eds.), Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 365-386.

Sites Jr, J.W., Reeder, T.W., Wiens, J.J., 2011. Phylogenetic insights on evolutionary novelties in lizards and snakes: sex, birth, bodies, niches, and venom. Annu Rev Ecol Evol S 42, 227-244.

Sneath, P.H., Sokal, R.R., 1973. Numerical taxonomy. The principles and practice of numerical classification.

Stephens, M., Smith, N.J., Donnelly, P., 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68, 978-989.

Swofford, D.L., 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, Massachusetts.

Tamura, K., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10, 512 - 526.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28, 2731-2739.

Taylor, E.H., 1942. Mexican snakes of the genera *Adelophis* and *Storeria*. Herpetologica 2, 75-79.

Townsend, J.P., 2007. Profiling phylogenetic informativeness. Syst Biol 56, 222-231.

Trapido, H., 1944. The snakes of the genus *Storeria*. Am Midl Nat 31, 1-84.

Varkey, A., 1979. Comparative cranial myology of North American natricine snakes. Milwaukee Public Museum Press, Milwaukee.

Vincent, S.E., Brandley, M.C., Herrel, A., Alfaro, M.E., 2009. Convergence in trophic morphology and feeding performance among piscivorous natricine snakes. J Evol Biol 22, 1203-1211.

Vincent, S.E., Dang, P.D., Herrel, A., Kley, N.J., 2006. Morphological integration and adaptation in the snake feeding system: a comparative phylogenetic study. J Evol Biol 19, 1545-1554.

Wiens, J.J. 2007. Species delimitation: new approaches for discovering diversity. Syst Biol 56, 875-878.

Wood, D.A., Vandergast, A.G., Lemos Espinal, J.A., Fisher, R.N., Holycross, A.T., 2011. Refugial isolation and divergence in the Narrowheaded Gartersnake species complex (*Thamnophis rufipunctatus*) as revealed by multilocus DNA sequence data. Mol Ecol 20, 3856-3878.

Xie, W., Lewis, P.O., Fan, Y., Kuo, L., Chen, M.H., 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst Biol 60, 150-160.

Yang, Z., 2006. Computational molecular evolution. Oxford University Press, Oxford.

APPENDIX A. SUPPLEMENTAL MATERIALS FOR CHAPTER 2

Figure A.1. Bayesian maximum clade credibility tree obtained from BEAST. Taxonomy key: NCYC= *Nerodia cyclopion*; NERY = *N. erythrogaster*; NFASC = *N. fasciata*; NFLO = *N. floridana*; NRHO = *N. rhombifer*; NSIP = *N. sipedon*; NTAX = *N. taxispilota*; RGRA = *Regina grahamii*; TROP = *Tropidoclonion.*

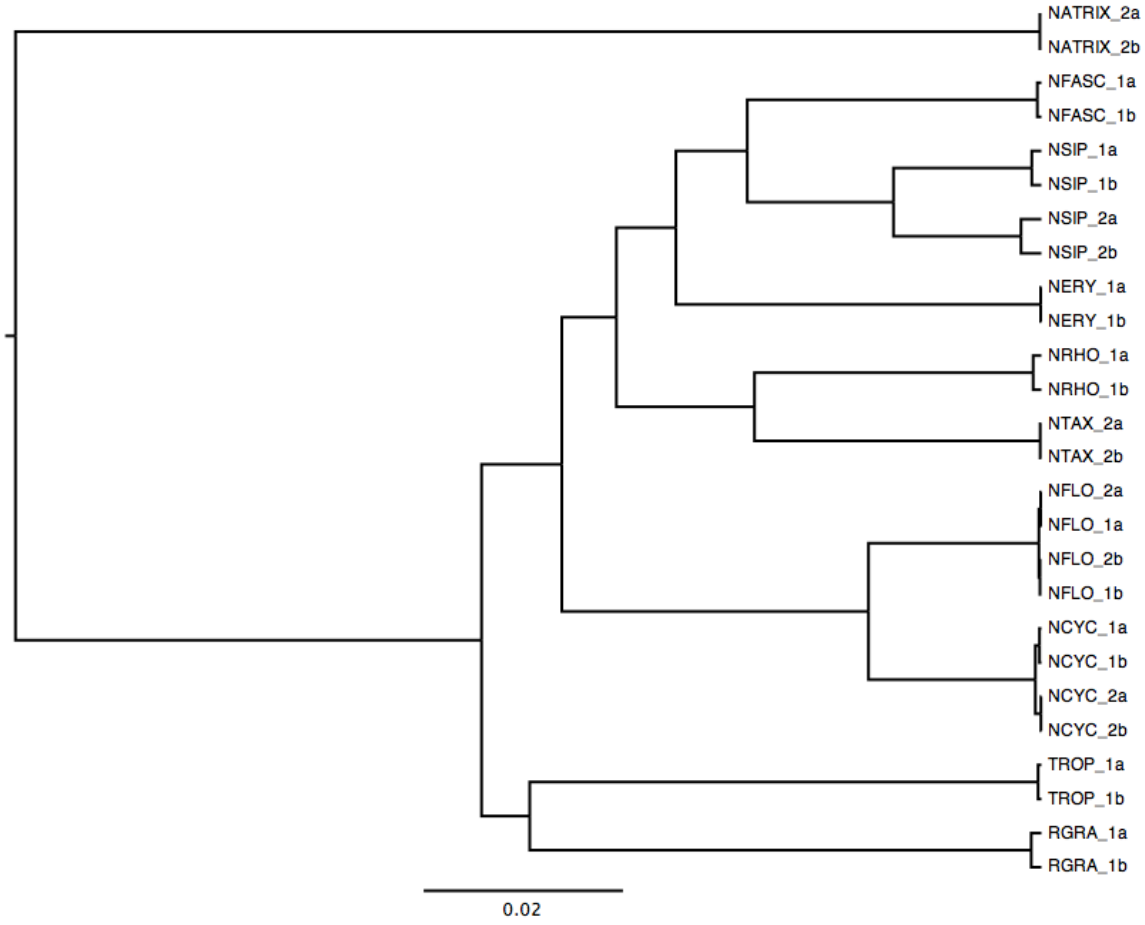
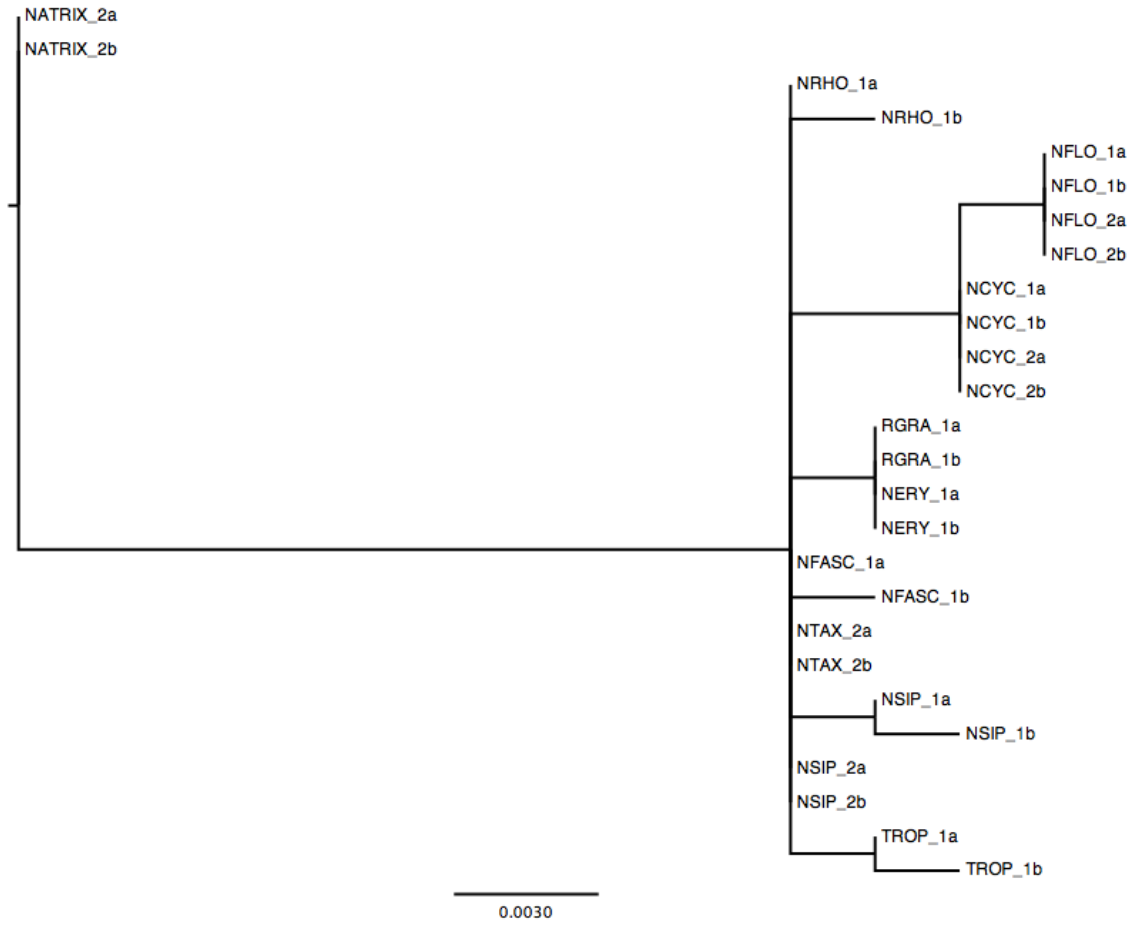
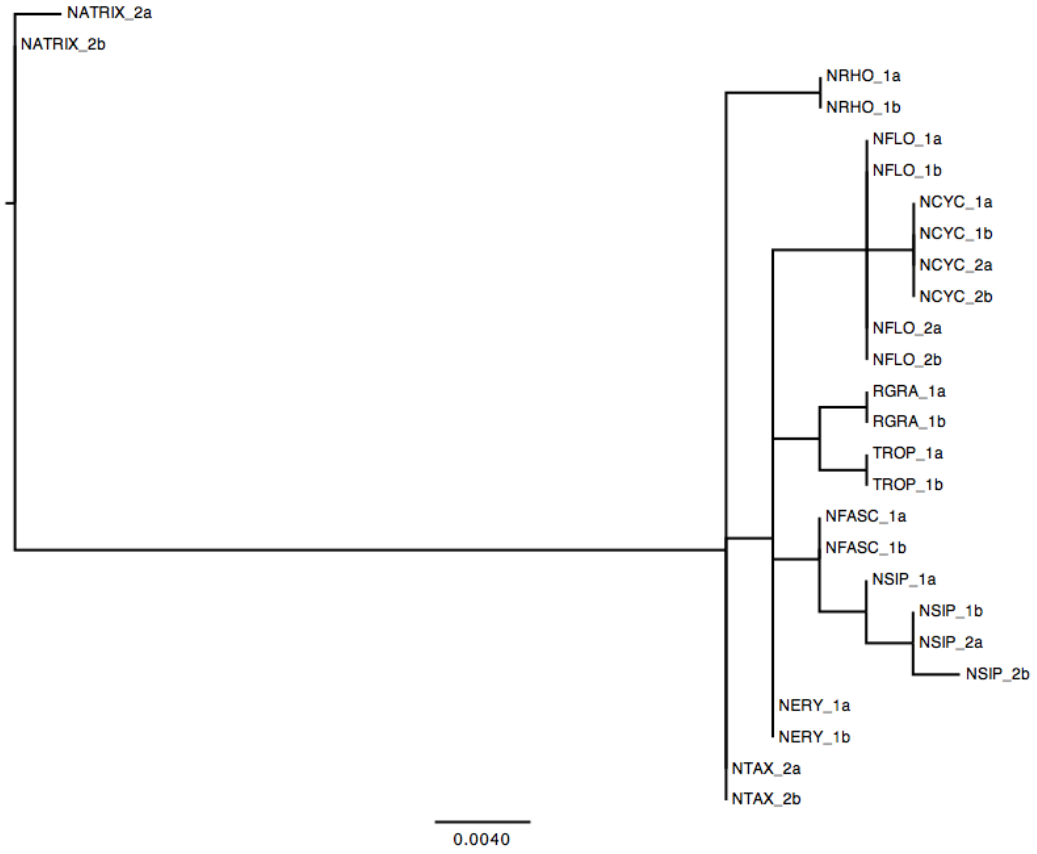


Figure A.2. Maximum likelihood gene trees. a) BDNF b) FSHR c) MC1R d) mtDNA (CYTB + ND4) e) NTF3 f) R35. Taxonomy key: NCYC= *Nerodia cyclopion*; NERY = *N. erythrogaster*; NFASC = *N. fasciata*; NFLO = *N. floridana*; NRHO = *N. rhombifer*; NSIP = *N. sipedon*; NTAX = *N. taxispilota*; RGRA = *Regina grahamii*; TROP = *Tropidoclonion.*

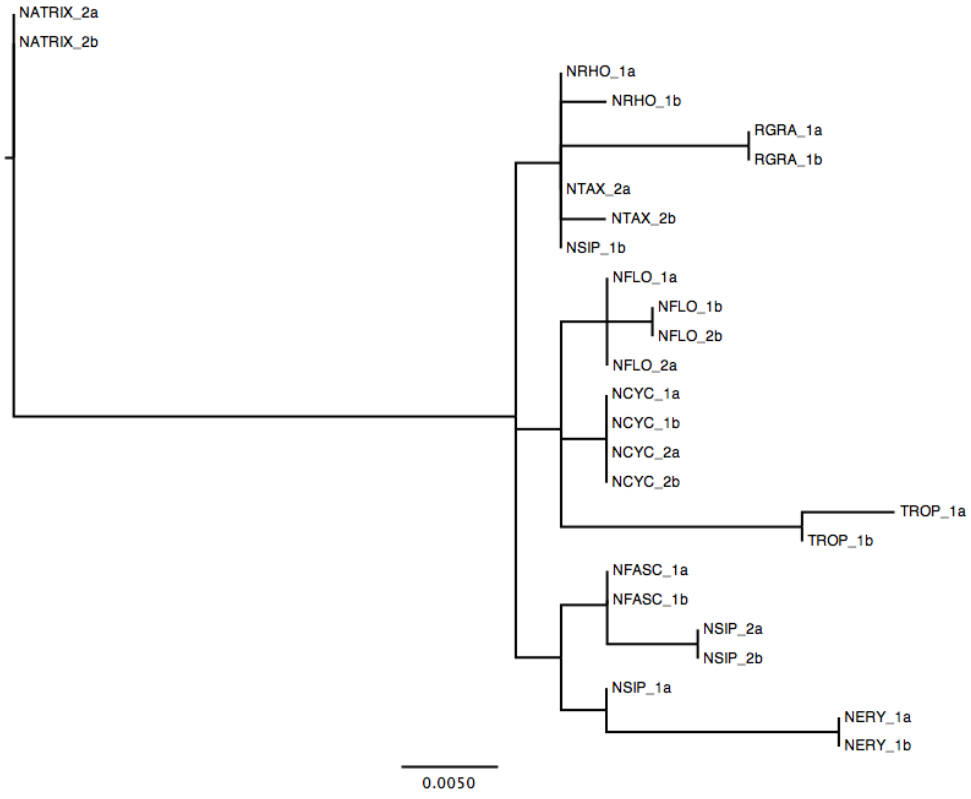
a)



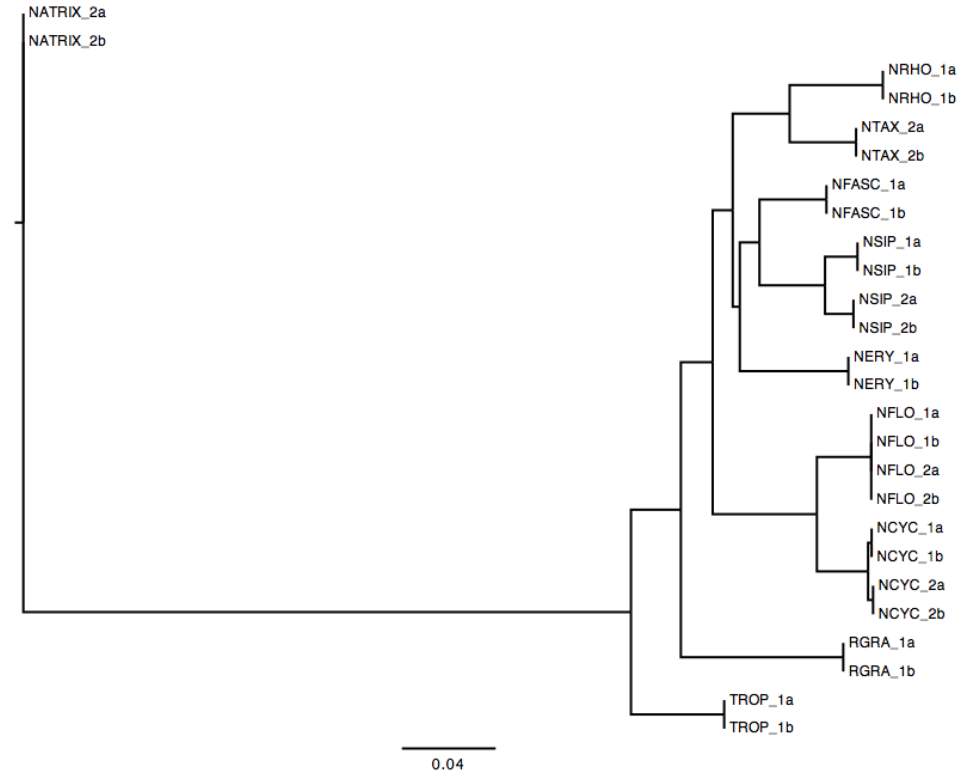
b)


NATRIX_2a
NATRIX_2b
NRHO_1a
NRHO_1b
NFLO_1a
NFLO_1b
NCYC_1a
NCYC_1b
NCYC_2a
NCYC_2b
NFLO_2a
NFLO_2b
RGRA_1a
RGRA_1b
TROP_1a
TROP_1b
NFASC_1a
NFASC_1b
NSIP_1a
NSIP_1b
NSIP_2a
NSIP_2b
NERY_1a
NERY_1b
NTAX_2a
NTAX_2b
0.0040

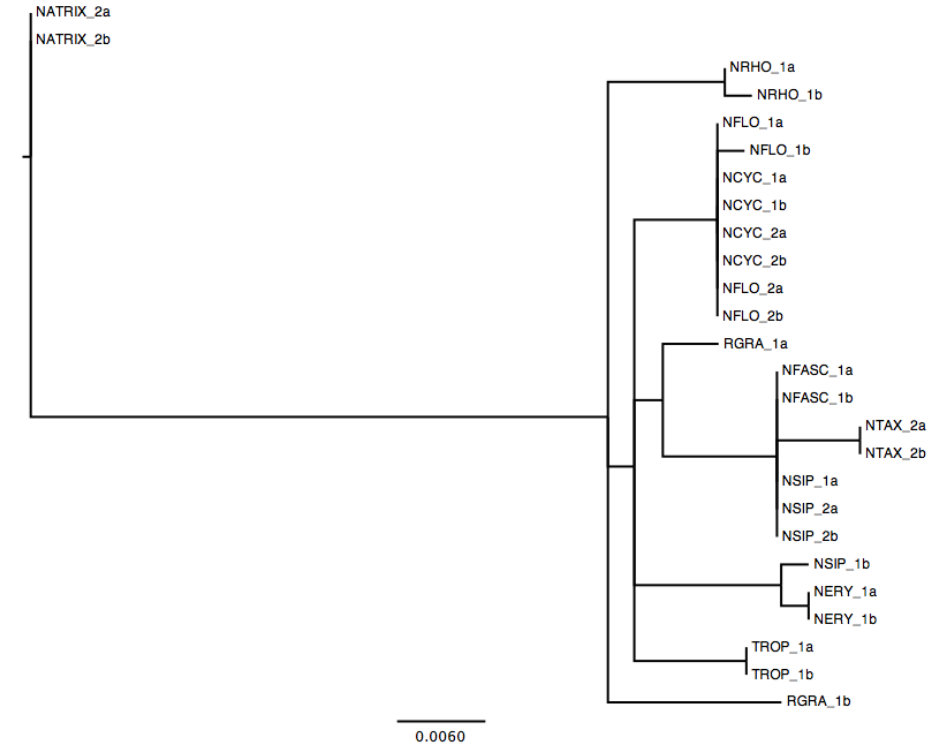
c)


NATRIX_2a
NATRIX_2b
NRHO_1a
NRHO_1b
RGRA_1a
RGRA_1b
NTAX_2a
NTAX_2b
NSIP_1b
NFLO_1a
NFLO_1b
NFLO_2b
NFLO_2a
NCYC_1a
NCYC_1b
NCYC_2a
NCYC_2b
TROP_1a
TROP_1b
NFASC_1a
NFASC_1b
NSIP_2a
NSIP_2b
NSIP_1a
NERY_1a
NERY_1b
0.0050

d)


NATRIX_2a
NATRIX_2b
NRHO_1a
NRHO_1b
NTAX_2a
NTAX_2b
NFASC_1a
NFASC_1b
NSIP_1a
NSIP_1b
NSIP_2a
NSIP_2b
NERY_1a
NERY_1b
NFLO_1a
NFLO_1b
NFLO_2a
NFLO_2b
NCYC_1a
NCYC_1b
NCYC_2a
NCYC_2b
RGRA_1a
RGRA_1b
TROP_1a
TROP_1b
0.04

e)


NATRIX_2a
NATRIX_2b
NRHO_1a
NRHO_1b
NFLO_1a
NFLO_1b
NCYC_1a
NCYC_1b
NCYC_2a
NCYC_2b
NFLO_2a
NFLO_2b
RGRA_1a
NFASC_1a
NFASC_1b
NTAX_2a
NTAX_2b
NSIP_1a
NSIP_2a
NSIP_2b
NSIP_1b
NERY_1a
NERY_1b
TROP_1a
TROP_1b
RGRA_1b
0.0060

f)

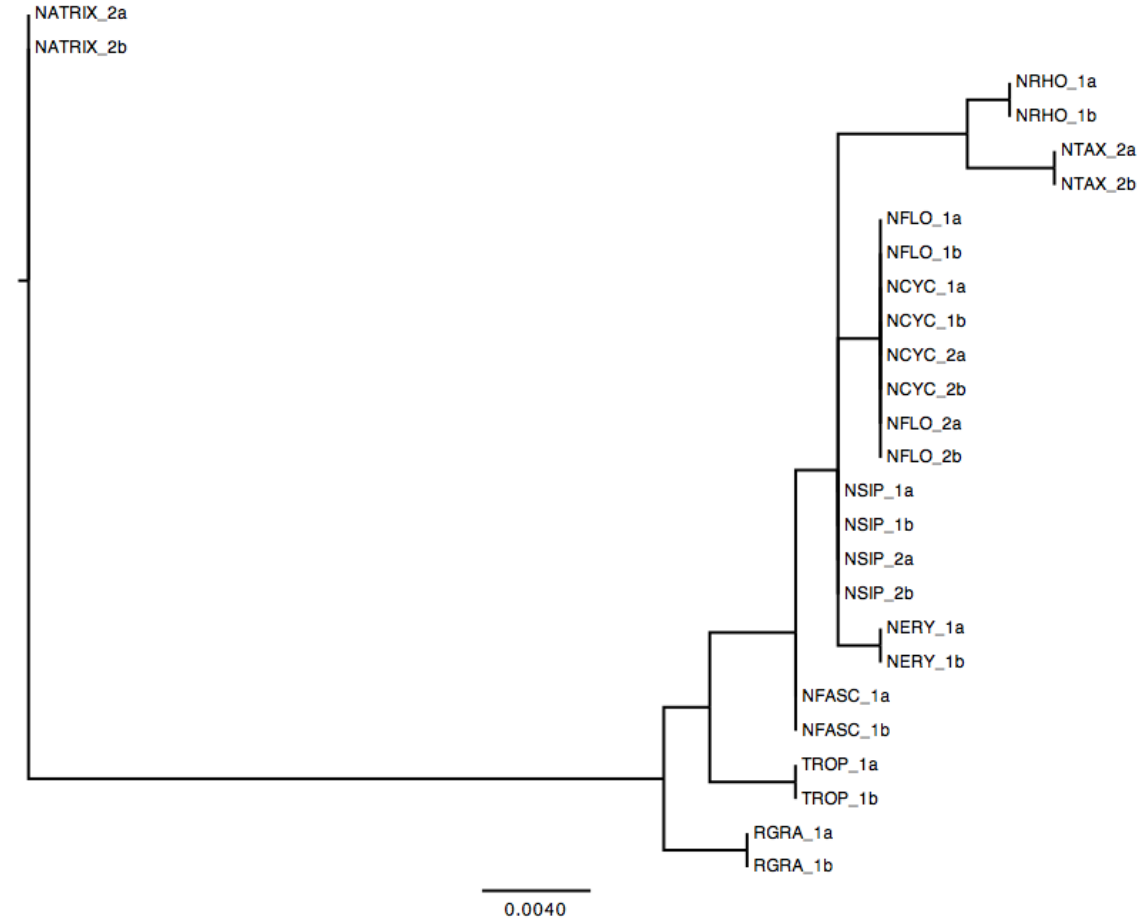
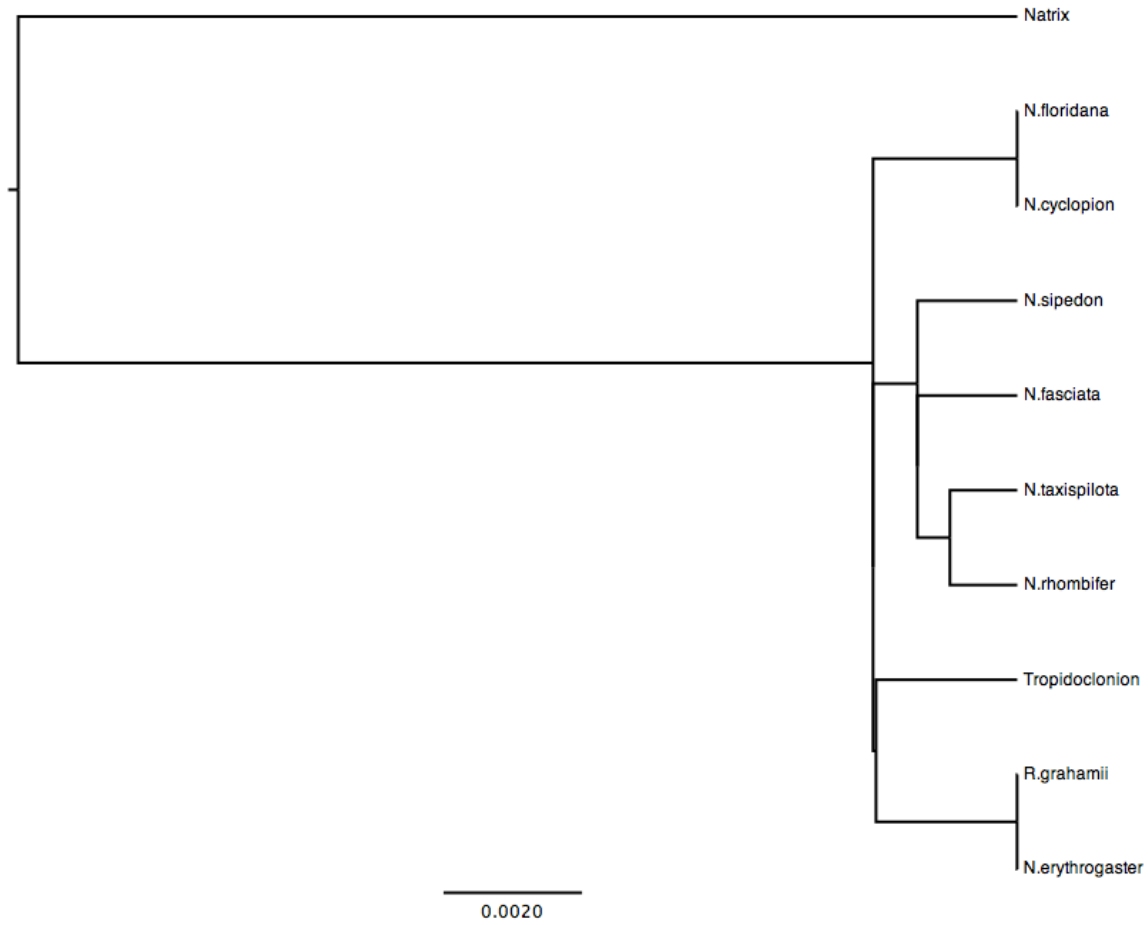
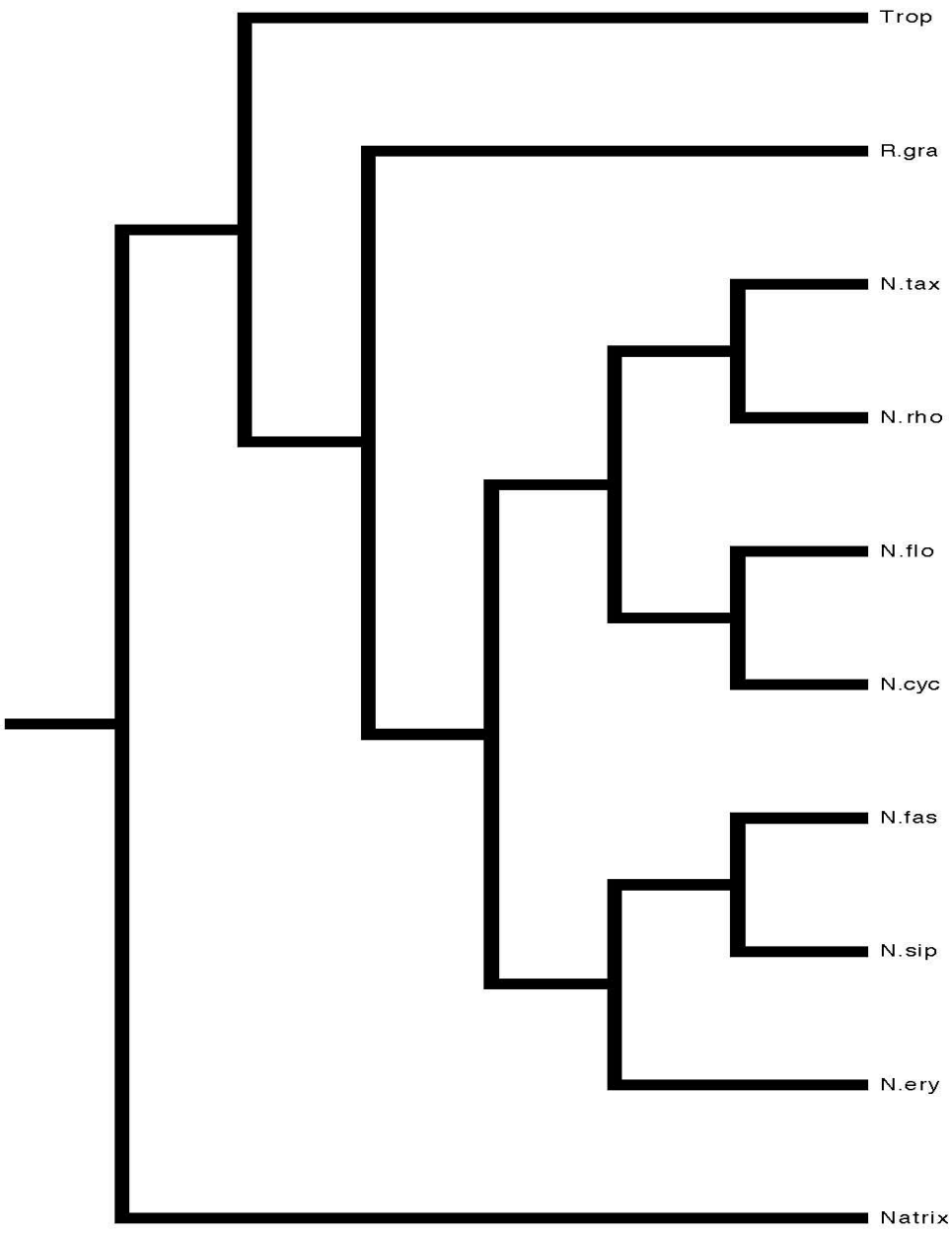

NATRIX_2a
NATRIX_2b
NRHO_1a
NRHO_1b
NTAX_2a
NTAX_2b
NFLO_1a
NFLO_1b
NCYC_1a
NCYC_1b
NCYC_2a
NCYC_2b
NFLO_2a
NFLO_2b
NSIP_1a
NSIP_1b
NSIP_2a
NSIP_2b
NERY_1a
NERY_1b
NFASC_1a
NFASC_1b
TROP_1a
TROP_1b
RGRA_1a
RGRA_1b
0.0040

Figure A.3. Species tree estimates obtained from a) STEM and b) Mesquite (MDC).

a)



b)



APPENDIX B. SUPPLEMENTAL MATERIALS FOR CHAPTER 3

Table B.1. Individuals representing the nine genera of Thamnophiini and *Natrix*, an Old World outgroup.

Scientific name	LSUMZ No.	Country	State
Adelophis foxi	H08272	México	Durango
Clonophis kirtlandii	H08189	USA	Illinois
Natrix natrix	H05128	na	na
Nerodia cyclopion	H08217	USA	Florida
Nerodia fasciata	H08208	USA	Alabama
Regina alleni	H08565	USA	Florida
Regina grahamii	H08235	USA	Louisiana
Regina rigida	H20703	USA	Louisiana
Regina septemvittata	H08166	USA	na
Seminatrix pygaea	H08993	USA	Georgia
Storeria dekayi	H07700	USA	Pennsylvania
Storeria storerioides	H08265	México	México
Thamnophis butleri	H07718	Canada	Ontario
Thamnophis proximus	H19835	USA	Louisiana
Tropidoclonion lineatum	H13044	USA	na
Virginia striatula	H18238	USA	Louisiana
Virginia valeriae	H16063	USA	Louisiana

Table B.2. Primers and sources of gene fragments used in this study. Shown for each gene are the primer sequence (in 5' to 3' orientation) and the source of each primer.

Gene		Oligo (5'-3')	Reference
BDNF	F	GACCATCCTTTTCCTKACTATGGTTATTTCATACTT	Leache and McGuire (2006)
	R	CTATCTTCCCCTTTTAATGGTCAGTGTACAAAC	
FSHR	F	CCDGATGCCTTCAACCCVTGTGA	Wiens et al. (2008)
	R	CCRAAYTTRCTYAGYARRATGA	
CYTB	F	TGATCTGAAAAACCACCGTTGTA	Alfaro and Arnold (2001)
	R	AATGGGATTTTGTCAATGTCTGA	
MC1R	F	TCAGCAACGTGGTGGA	Austin et al. (2009)
	R	ATGAGGTAGAGGCTGAAGTA	
ND4	F	TGACTACCAAAAGCTCATGTAGAAGC	Forstner et al. (1995)
	R	TTTTACTTGGATTTGCACCA	Skinner et al. (2006)
NT3	F	ATGTCCATCTTGTTTTATGTGATATTT	Wiens et al. (2008)
	R	ACRAGTTTRTTGTTYTCTGAAGTC	
R35	F	TCTAAGTGTGGATGATYTGAT	Fry et al. (2006)
	R	CATCATTGGRAGCCAAAGAA	

Table B.3. Descriptive statistics of sequenced loci. Shown for each gene are the length of sequence (bp), the number of variable sites (s), the model of sequence evolution (model), and p-value of the z-factor-based test of purifying selection.

Gene	bp	s	model	dS>dN*
BDNF	557	23	K80 + G	0
FSHR	511	31	HKY + G	0
CYTB	521	180	HKY + G	0
MC1R	435	35	HKY + I	0.005
ND4	614	237	HKY + G	0
NT3	561	64	K80 + G	0
R35	645	53	HKY + G	0.01
Total	3844	623	N/A	N/A

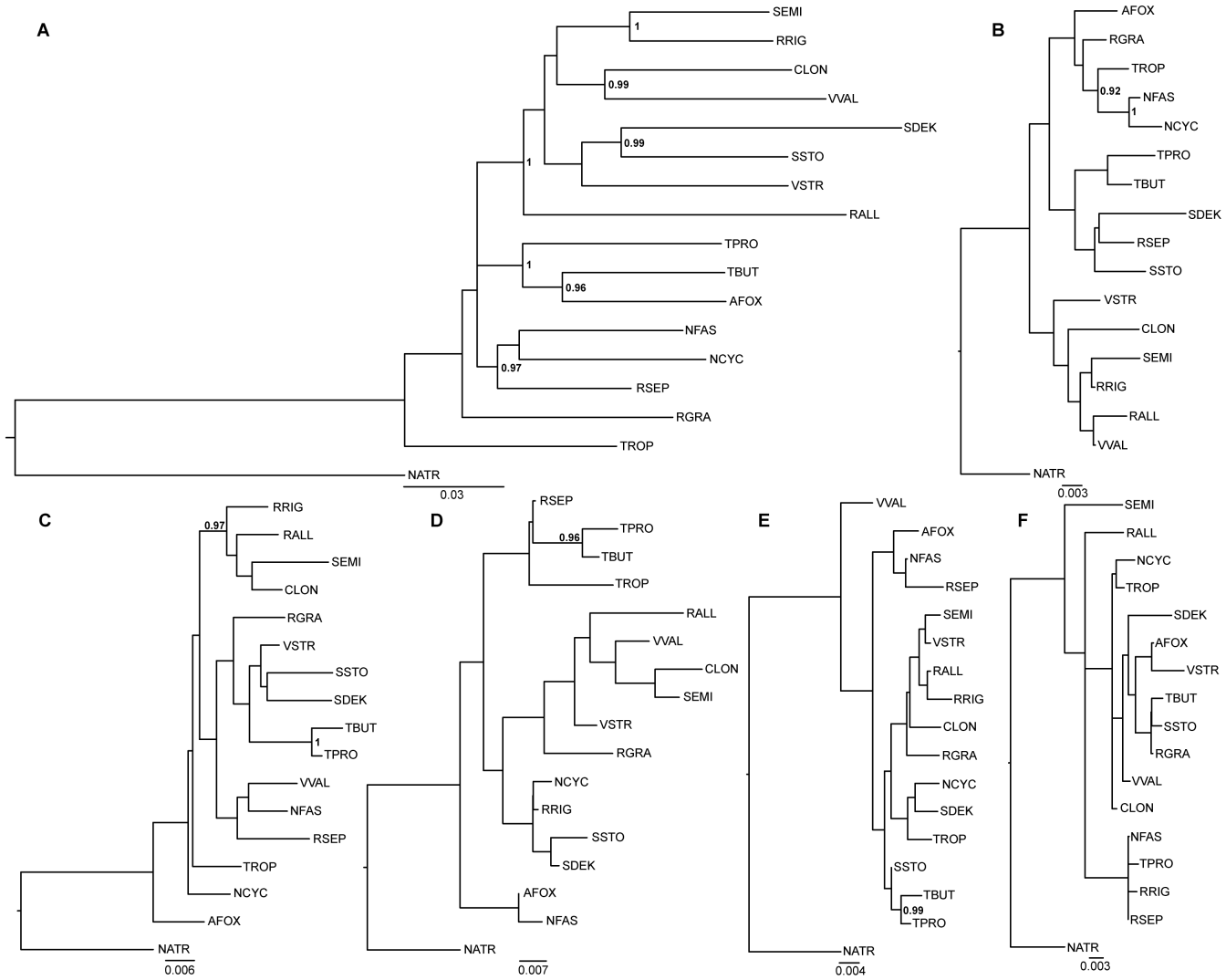
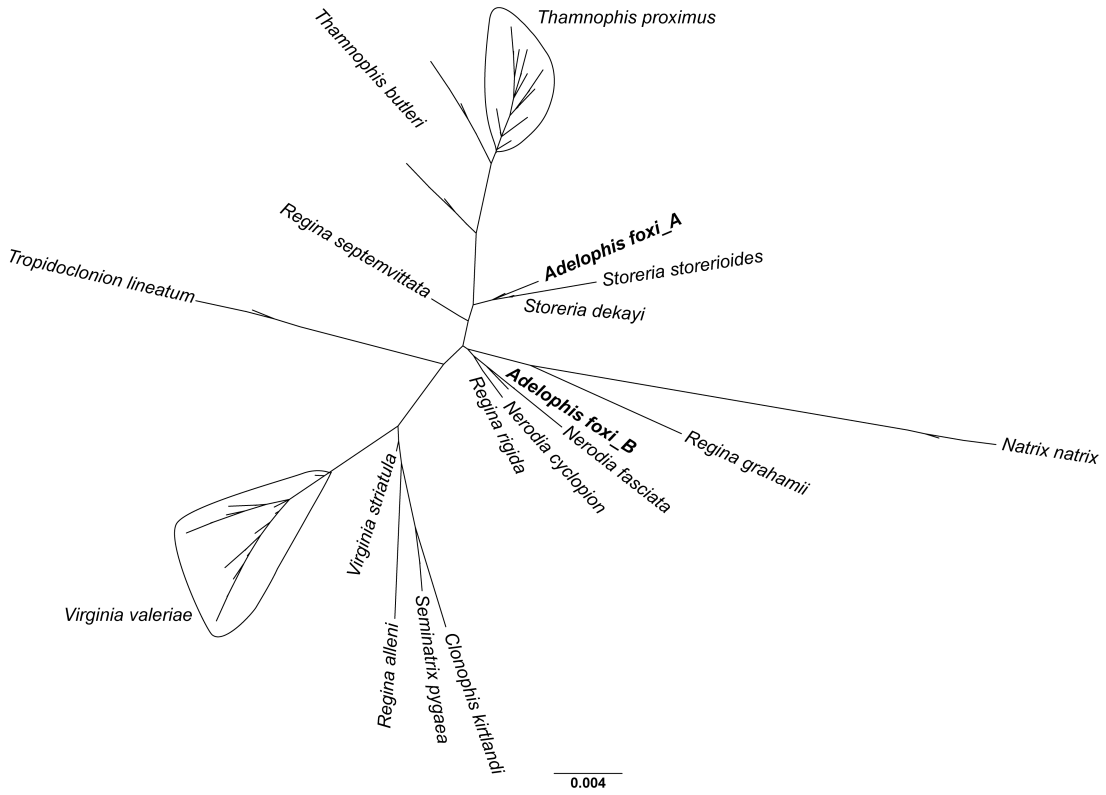


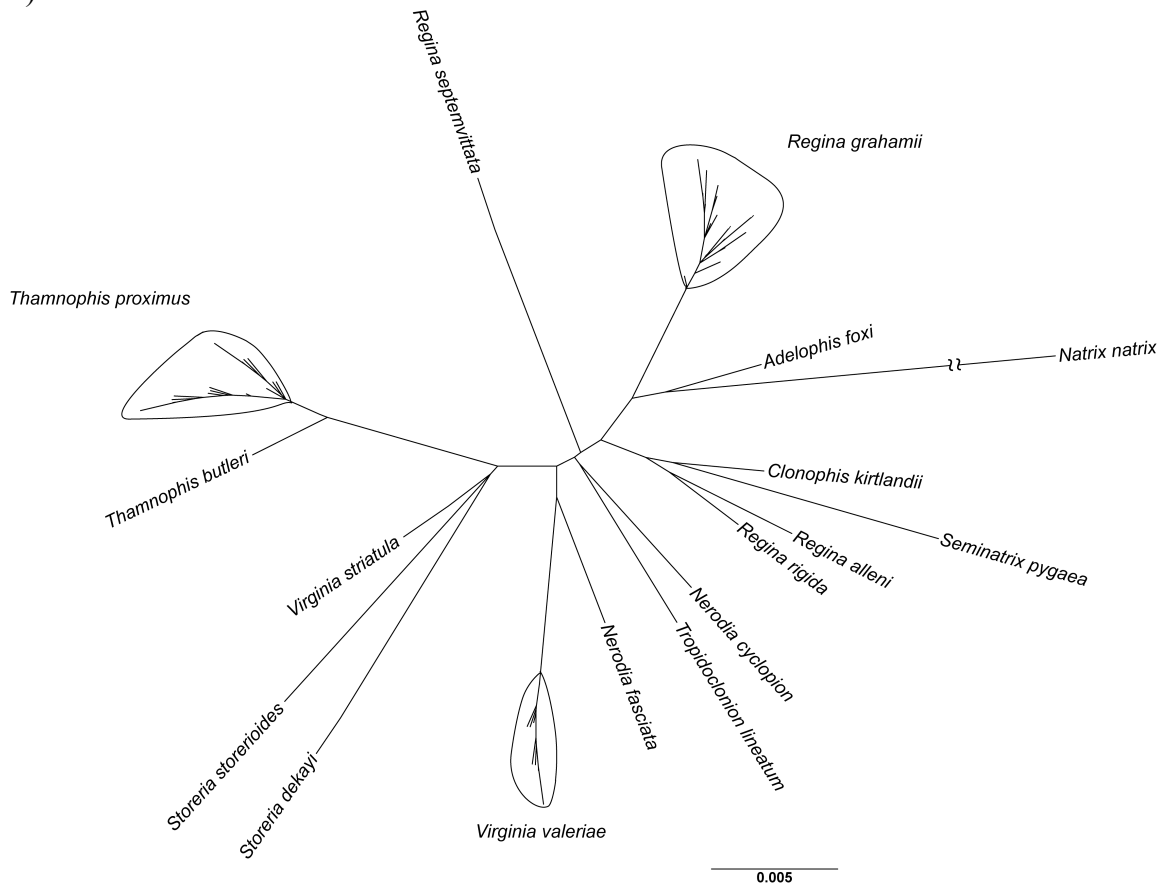
Figure B.1. Maximum clade credibility Bayesian gene trees. A) Mitochondrial data (CYTB + ND4), B) R35, C) NT3, D) MC1R, E) FSHR, F) BDNF. Unlabeled nodes were not supported with greater than 0.9 Bayesian posterior probability.

Figure B.2. Neighbor-joining estimates for all possible phase reconstructions for a) MC1R and b) NT3. Circled regions encompass all possible phases for ambiguous individuals, except for *Adelophis foxi* in MC1R, with both potential phases shown in bold.

A)



B)



Appendix B References

Austin, C.C., Spataro, M., Peterson, S., Jordan, J., Mcvay, J.D., 2010. Conservation genetics of Boelen's python (Morelia boeleni) from New Guinea: reduced genetic diversity and divergence of captive and wild animals. Cons Genet 11, 889-896.

Forstner, M.R., Davis, S.K., Arevalo, E., 1995. Support for the hypothesis of anguimorph ancestry for the suborder Serpentes from phylogenetic analysis of mitochondrial DNA sequences. Mol Phylogenet Evol 4, 93-102.

Fry, B.G., Vidal, N., Norman, J.A., Vonk, F.J., Scheib, H., Ramjan, S.F.R., Kuruppu, S., Fung, K., Hedges, S.B., Richardson, M.K., Hodgson, W.C., Ignjatovic, V., Summerhayes, R., Kochva, E., 2006. Early evolution of the venom system in lizards and snakes. Nature 439, 584-588.

Leache, A.D., McGuire, J.A., 2006. Phylogenetic relationships of horned lizards (*Phrynosoma*) based on nuclear and mitochondrial data: Evidence for a misleading mitochondrial gene tree. Mol Phylogenet Evo 39, 628-644.

Skinner, A., Donnellan, S.C., Hutchinson, M.N., Hutchinson, R.G., 2005. A phylogenetic analysis of *Pseudonaja* (Hydrophiinae, Elapidae, Serpentes) based on mitochondrial DNA sequences. Mol Phylogenet Evol 37, 558-571.

Wiens, J.J., Kuczynski, C.A., Smith, S.A., Mulcahy, D.G., Sites, J.W., Jr., Townsend, T.M., Reeder, T.W., 2008. Branch lengths, support, and congruence: testing the phylogenomic approach with 20 nuclear loci in snakes. Syst Biol 57, 420-431.

APPENDIX C. SUPPLEMENTAL MATERIALS FOR CHAPTER 4

Table C.1. Materials examined.

Scientific name	Specimen No.	Country	State
Adelophis foxi	LSUMZ 40846	México	Durango
Clonophis kirtlandii	LSUMZ 39566	USA	Illinois
Natrix natrix	H05128	na	na
Nerodia clarkii	LSUMZ 43426	USA	Alabama
Nerodia cyclopion	JDM 1034	USA	Louisiana
Nerodia erythrogaster	JDM 1004	USA	Texas
Nerodia fasciata	LSUMZ 40040	USA	Alabama
Nerodia floridana	LSUMZ 40090	USA	Florida
Nerodia rhombifer	H21296	USA	Louisiana
Nerodia sipedon	LSUMZ 40906	USA	Georgia
Nerodia taxispilota	LSUMZ 40308	USA	Florida
Regina alleni	LSUMZ 40570	USA	Florida
Regina grahamii	LSUMZ 40330	USA	Louisiana
Regina rigida	LSUMZ 40503	USA	Louisiana
Regina septemvittata	LSUMZ 40101	USA	na
Seminatrix pygaea	LSUMZ 42686	USA	Georgia
Storeria dekayi	LSUMZ 39878	USA	Pennsylvania
Storeria hidalgoensis	JAC 23435	México	Jalisco
Storeria occipitomaculata	LSUMZ 80971	USA	Louisiana
Storeria storerioides	LSUMZ 40790	México	México
Thamnophis atratus	LSUMZ 44386	USA	California
Thamnophis brachystoma	LSUMZ 58447	USA	Pennsylvania
Thamnophis butleri	LSUMZ 39656	Canada	Ontario
Thamnophis chrysocephalus[1]	HCD7310	México	na
Thamnophis couchii	H08146	na	na
Thamnophis cyrtopsis	LSUMZ 40426	USA	New Mexico
Thamnophis elegans	LSUMZ 39641	USA	New Mexico
Thamnophis eques	LSUMZ 40752	México	Durango
Thamnophis errans[2]	LSUMZ 16999	México	Durango
Thamnophis fulvus	LSUMZ 57127	Guatemala	Jalapa
Thamnophis lineri	JAC 21406	México	Oaxaca
Thamnophis bogerti	JAC 21416	México	Oaxaca
Thamnophis conanti	JAC 22810	México	Puebla
Thamnophis couchii	LSUMZ 37179	USA	California
Thamnophis marcianus	LSUMZ 48745	USA	Texas
Thamnophis melanogaster	LSUMZ 37429	México	Michoacán
Thamnophis nigronuchalis	LSUMZ 40849	México	Durango
Thamnophis ordinoides	LSUMZ 40130	Canada	British Columbia

Table C.1 continued.

Scientific name	Specimen No.	Country	State
Thamnophis proximus	LSUMZ 87348	USA	Louisiana
Thamnophis pulchrilatus	LSUMZ 35379	México	Durango
Thamnophis radix	H02935	USA	Wisconsin
Thamnophis rufipunctatus	LSUMZ 40538	na	na
Thamnophis sauritus	LSUMZ 41508	USA	Florida
Thamnophis scalaris	LSUMZ 42639	México	México
Thamnophis scalaris	LSUMZ 42641	México	México
Thamnophis scaliger	LSUMZ 42640	México	México
Thamnophis sirtalis	LSUMZ 41181	USA	Maine
Thamnophis sumichrasti[3]	LSUMZ 11114	México	Hidalgo
Thamnophis validus	JRM 4541	México	Sinaloa
Tropidoclonion lineatum	H13044	USA	na
Virginia striatula	LSUMZ 83481	USA	Louisiana
Virginia valeriae	LSUMZ 81173	USA	Louisiana

Collection abbreviations: LSUMZ = Louisiana State University Museum of Zoology; H = LSUMZ tissue catalog; JAC = Jonathan A. Campbell field catalog; JDM = John D. McVay field catalog; JRM = Joseph R. Mendelson III field catalog. Where noted, the ND4 sequence data was taken from genbank: [1]AF420098; [2]EF417363; [3]AF420200.

Table C.2. Oldest discovered fossils of thamnophiine snakes.

Genus	Species	Oldest Period	Age range (MYA)
Neonatrix	*elongata*†	Hemingfordian	20.6-16.3
Neonatrix	*intera*†	Early Barstovian	16.3-13.6
Neonatrix	*magna*†	Medial Barstovian	16.3-13.6
Nerodia		Medial Barstovian	16.3-13.6
Nerodia	*erythrogaster*	Irvingtonian I	1.9-0.9
Nerodia	*fasciata*	Irvingtonian I	1.9-0.9
Nerodia	*rhombifer*	Blancan V	2.6-1.9
Nerodia	*floridana*	Irvingtonian I	1.9-0.9
Nerodia	*sipedon*	Blancan V	2.6-1.9
Nerodia	*taxispilota*	Rancholabrean II	0.15-0.01
Nerodia	*hibbardi*†	Blancan III	3.7-3.2
Nerodia	*hillmani*†	Clarendonian I	13.6-10.3
Regina	*sp.*	Irvingtonian I	1.9-0.9
Regina	*alleni*	Irvingtonian I	1.9-0.9
Regina	*grahamii*	Blancan V	2.6-1.9
Regina	*intermedia*†	Irvingtonian I	1.9-0.9
Regina	*septemvittata*	Rancholabrean II	0.15-0.01
Storeria	*sp.*	Irvingtonian II	0.9-0.4
Storeria	*cf. dekayi*	Rancholabrean I	0.4-0.15
Storeria	*dekayi*	Rancholabrean II	0.15-0.01
Storeria	*occipitomacula*	Rancholabrean II	0.15-0.01
Thamnophis		Medial Barstovian	16.3-13.6
Thamnophis	*brachystoma*	Rancholabrean II	0.15-0.01
Thamnophis	*couchii*	Rancholabrean II	0.15-0.01
Thamnophis	*cf. cyrtopsis*	Rancholabrean II	0.15-0.01
Thamnophis	*elegans*	Irvingtonian II	0.9-0.4
Thamnophis	*marcianus*	Blancan III	3.7-3.2
Thamnophis	*proximus*	Blancan V	2.6-1.9
Thamnophis	*cf. sirtalis*	Early Hemphillian	10.3-4.9
Thamnophis	*cf. sauritus*	Blancan IV	3.2-2.6
Thamnophis	*radix*	Blancan II	4.25-3.7
Thamnophis	*sirtalis*	Blancan II	4.25-3.7
Tropidoclonion	*lineatum*	Irvingtonian I	1.9-0.9
Virginia		Irvingtonian I	1.9-0.9
Virginia	*striatula*	Rancholabrean II	0.15-0.01
Virginia	*valeriae*	Rancholabrean II	0.15-0.01

Table C.3. Ecological data for thamnophiine snakes.

Genus	species	Diet	Habitat	Reference
Adelophis	*foxi*	?	terrestrial?	1
Adelophis	*copei*	?	near water	1
Clonophis	*kirtlandii*	slug/earthworm	terrestrial	2
Nerodia	*clarkii*	fish	aquatic	3
Nerodia	*cyclopion*	fish	aquatic	3
Nerodia	*harteri*	fish	aquatic	3
Nerodia	*fasciata*	generalist	aquatic	3
Nerodia	*sipedon*	generalist	aquatic	3
Nerodia	*erythrogaster*	generalist	aquatic	3
Nerodia	*floridana*	amphibians/fish	aquatic	3
Nerodia	*taxispilota*	fish	aquatic	3
Nerodia	*rhombifer*	fish	aquatic	3
Regina	*alleni*	durophagy	aquatic	3
Regina	*grahami*	stenophagy	aquatic	3
Regina	*rigida*	durophagy	aquatic	3
Regina	*septemvittata*	stenophagy	aquatic	3
Seminatrix	*pygaea*	generalist	aquatic	3
Storeria	*dekayi*	slug/earthworm/snails	semi-fossorial	2
Storeria	*occipitomacula*	slug/earthworm/snails	semi-fossorial	2
Storeria	*hidalgoensis*	?	semi-fossorial	4
Storeria	*storerioides*	?	semi-fossorial	4
Thamnophis	*atratus*	amphibians/fish	semiaquatic	5
Thamnophis	*butleri*	slug/earthworm	near water	5
Thamnophis	*brachystoma*	slug/earthworm	near water	5
Thamnophis	*radix*	generalist	near water	5
Thamnophis	*elegans*	mixed pops	mixed	5
Thamnophis	*errans*	?	?	5
Thamnophis	*eques*	amphibians	near water	5
Thamnophis	*fulvus*	amphibians	near water	5
Thamnophis	*pulchrilatus*	?	?	5
Thamnophis	*cyrtopsis*	amphibians	near water	5
Thamnophis	*proximus*	amphibians	near water	5
Thamnophis	*sirtalis*	mixed pops	mixed	5
Thamnophis	*sauritus*	amphibians/fish	near water	5
Thamnophis	*scaliger*	amphibians	near water	5
Thamnophis	*scalaris*	lizards?	terrestrial	5
Thamnophis	*melanogaster*	mixed pops	aquatic	5
Thamnophis	*sumichrasti*	?	?	5
Thamnophis	*chrysocephalus*	?	?	5
Thamnophis	*couchii*	fish/amphibians	aquatic	5
Thamnophis	*hammondi*	amphibians	aquatic	5
Thamnophis	*godmani*	mouse?	near water	5
Thamnophis	*validus*	fish/amphibians	aquatic	5
Thamnophis	*rufipunctatus*	fish	aquatic	5
Thamnophis	*nigronuchalis*	fish	aquatic	5

Table C.3 continued.

Genus	species	Diet	Habitat	Reference
Thamnophis	*marcianus*	mixed pops	mixed	5
Thamnophis	*ordinoides*	slug/earthworm	terrestrial	5
Thamnophis	*mendax*	salamanders?	terrestrial	5
Thamnophis	*gigas*	fish/amphibians	aquatic	5
Thamnophis	*exsul*	?	?	5
Thamnophis	*rossmani*	fish	?	5
Thamnophis	*lineri*	?	near water	5

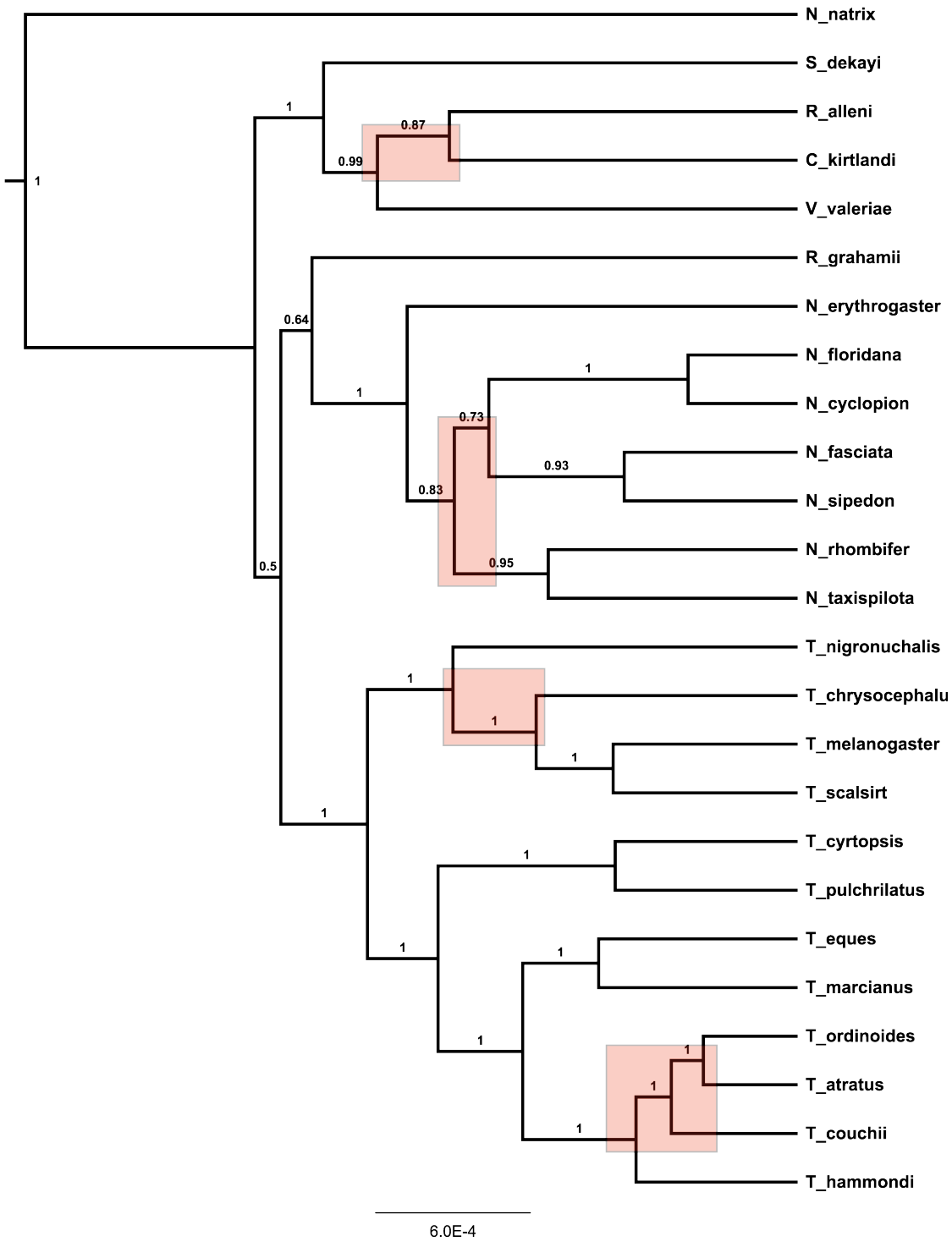


Figure C.1. Multi-locus Bayesian concatenated phylogeny estimate for individuals represented by all loci used in the dataset. Nodes highlighted red are areas where coalescent-based species tree conflicted (though not statistically supported) in topology.

APPENDIX D. DOUBLE-DIGEST ILLUMINA LIBRARY PREPARATION

The objective of this to use a double restriction enzyme digest to reduce the representation of genomic data across three species of *Nerodia*; the resulting data will be used for population genetic studies across a putative hybrid matrix between three species. Samples will be indexed and multiplexed on a single lane of an Illumina GAIIx flow cell.

Background: The indexing method is adapted from Meyer and Kircher (2010); the initial library preparation is modified from an amplified fragment length polymorphism (AFLP) protocol. Our method largely follows the protocol from Meyer and Kircher, briefly outlined below, with key changes, also detailed below, to allow for reduced representation.

Step 1: restriction enzyme digestion of tDNA samples. Following manufacturers' protocols, digest (AMT) purified DNA using two restriction enzymes; one common and one rare. The rare enzyme is used to sufficiently reduce the representation of the genome to homologous fragments across individuals; the common enzyme is used to cut fragments into lengths appropriate for analysis in the Illumina GAIIx. Here we use MseI (common) and EcoRI (rare).

Table D.1. Reaction conditions for restriction enzyme digestion of DNA samples.

Reagent	Reaction Vol µl	Final Concentration
DNA	varies	20ng/uL
NEB Buffer 4 10X	5	1X
BSA 100X	1	2X (increased for double the enzymes)
EcoRI (10,000U/mL)	0.5uL (5U)	0.1U/uL
MseI (10,000U/mL)	0.5uL (5U)	0.1U/uL
H_2O	To 50uL reaction volume	

Add 1µg of DNA to reaction, based on estimation of concentration (Nanodrop); for example, if DNA concentration is 200ng/µL, add 5µL DNA and 38µL H2O (50 [total reaction volume] – 7 [total reagent volume] – 5[DNA] = 38). Incubate and inactivate restriction enzymes according to manufacturers' protocols. For used enzymes: 60 minutes at 37°C followed by 20 minutes at 65°C. For specific heat inactivation profiles of restriction enzymes see the manufacturer's website (http://www.neb.com/). Reaction should result in fragmented DNA appearing as a smear on agarose gel. This procedure replaces steps 3-5 in Meyer and Kircher. From this point the Meyer and Kircher protocol was followed through step 24 (skipping steps 19-21), after which we size-selected amplicons to further reduce representation of the samples. Fragments between 300-400bp were cut by hand from gels visualized with UV transillumination; samples were then purified using a QIAquick gel extraction kit (Qiagen, Germantown, MD). Following purification, samples were quantified using a Bioanalyzer (Agilent Technologies, Santa Clara, CA). Samples were normalized and pooled; amplification and sequencing were performed on a cBot and Illumina GAIIx (Illumina, San Diego, CA) following manufacturers' protocols.

After sequencing, samples Detection of homologous reads across individuals and single nucleotide polymorphisms (SNPS) were detected using Stacks (Catchen et al., 2011)

Output from the Illumina GAIIx is in fastq format, with quality of each base scaled by "Phred33" scores. To filter out low quality bases, we used a python script to change all bases with Q<15 to 'n'.

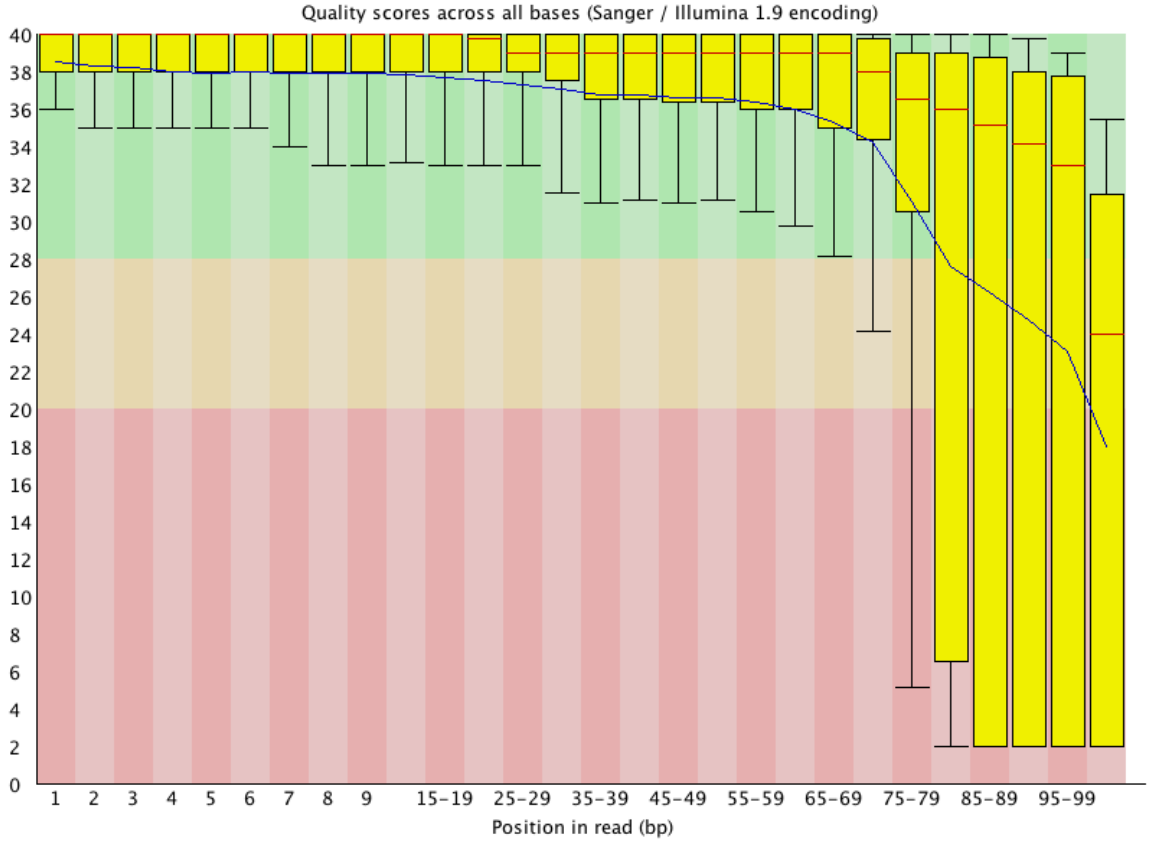


Figure D.1. Quality scores of reads from LSUMZ 44749, visualized with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Results from this individual are consistent with those from all other samples.

These data were initially analyzed with Stacks; however, initial results indicated that Stacks modified all 'n' bases into A's. As this may potentially bias our interpretation, we created two perl scripts to further filter the data prior to analysis. First, given the lower average quality of latter bases among reads, we trimmed each read to various lengths (50, 55, 60, 70 and 80 base pairs); a second perl script was then implemented to remove any read still containing an 'n'. The goal of this was to assess the trade-off between length of reads, number of reads, and number of catalogs (homologous reads across individuals). Expectedly, the number of accepted reads increased as the length of each read was shortened, truncating low-quality bases (Figure D.2.)

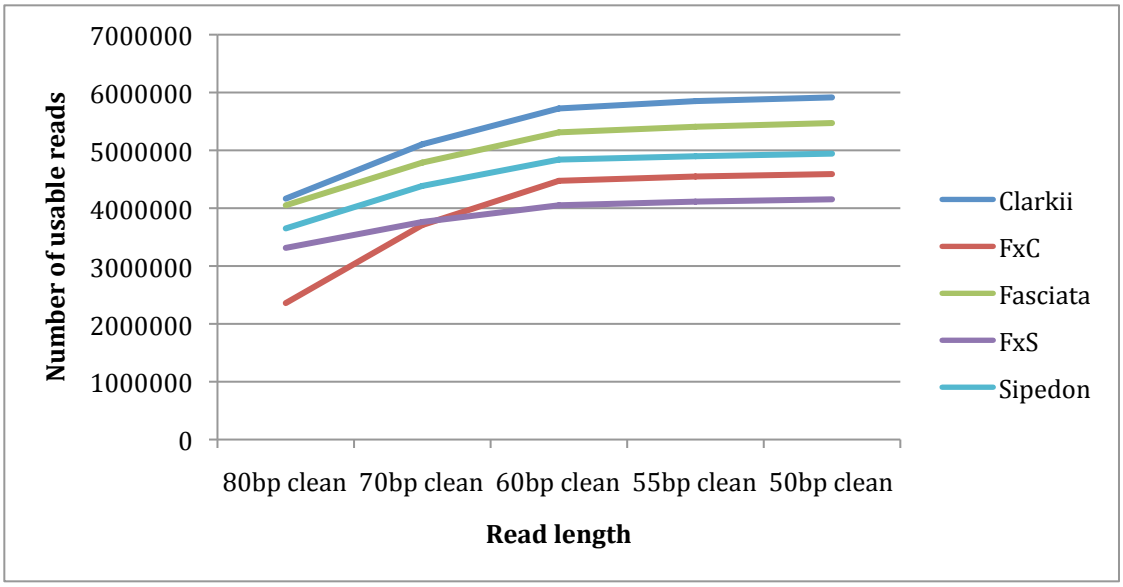


Figure D.2. Number of usable (no bases with Q<15) reads per individual by length of trimmed sequences.

For each read length, I utilized the *denovo_map* feature of Stacks, which compiles homologous reads within individuals (stacks), then searches for homology among all samples (catalogs). Because so few individuals were used, we sought to extrapolate the potential number of homologous regions available with this method by searching for catalogs that were found in all possible combinations of three and four individuals (Figure D.3.).

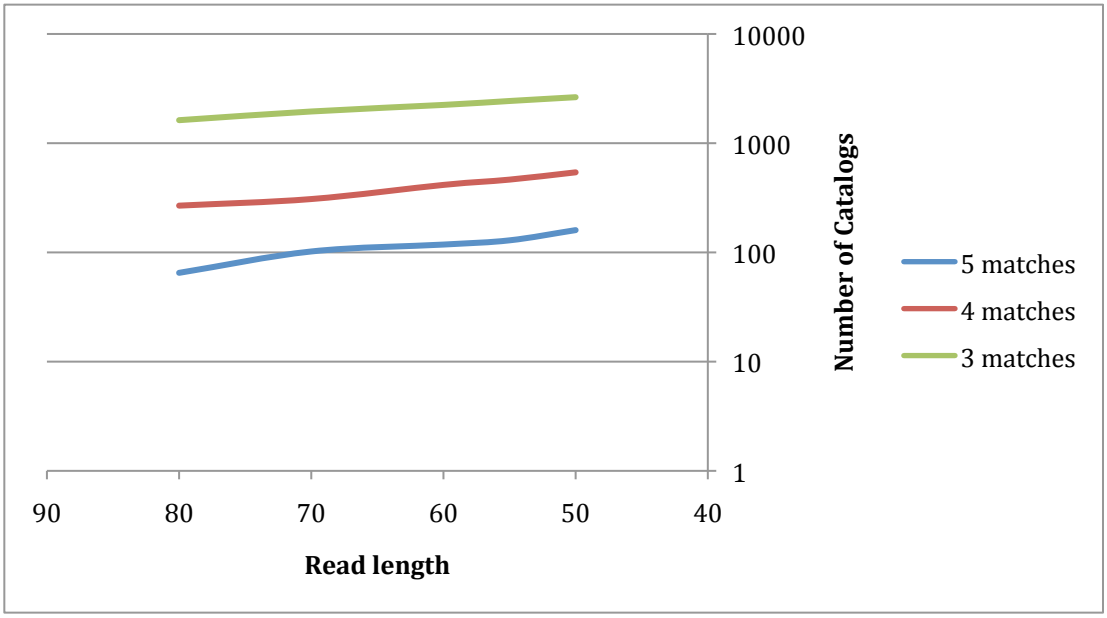


Figure D.3. The number of homologous regions sampled, given subsets of sequence length and individuals.

Our results suggest that a decrease in the number of catalogs due to a decrease in read length was overcome by the increase in number of usable reads as sequences were shortened. Moreover, the number of catalogs increased exponentially as we sub-sampled the individuals, indicating a potential to sample thousands of homologous loci.

Potential improvements.—There are a number of ways in which the quality of results may be improved upon, both in terms of quality of individual reads and in number of homologous loci sampled. A potential cause of both was an excess number of polymerase chain reaction (PCR) cycles performed during the protocol. Specifically, presence of heterodimers in the "bioanalyzed" product led me to perform additional 2-cycle PCRs to eliminate this effect. Additionally, multiple attempts at optimizing product amount across the protocol resulted in over 30 cycles of PCR performed; the goal of this step is not to amplify fragments, but rather to connect sequencing adapters to fragments. My mistakes are a potential cause of PCR-based sampling bias, which may lead to fewer homologous loci incorporated into the sequencing steps. Published methods similar to the above are now available ; use of these would minimize the need for optimization and troubleshooting. Additionally, precision of size selection could be improved using automated methods (e.g., Pippin Prep™).

Appendix D References

Catchen, J, A. Amores, P. Hohenlohe, W. Cresko, and J. Postlethwait. 2011. Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics, 1,171-182.

Meyer, M., and Kircher, M. 2010. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harbor Protocols, doi:10.1101/pdb.prot5448.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7, e37135.

VITA

John David McVay was born in Chapel Hill, North Carolina in 1977. At five days old, his family moved to Columbus, Georgia, then to Kiln, Mississippi and Baton Rouge, Louisiana before completing his childhood in Lubbock, Texas. John Graduated from Lubbock High School in 1995, where he was a member of the band and the varsity swim team.

John moved to Austin in 1995 to attend the University of Texas. As a sophomore, he enrolled in Herpetology through the grace of Eric Pianka (it was an upper-division course). Following this career-directing experience, he began to work in David Cannatella's laboratory under the tutelage of then graduate student Rafe Brown. His love of herpetology and systematics were strengthened through the many interactions with the faculty and graduate students in the Zoology Department.

After working as a fishmonger for two years, John decided to return to academia, seeking his masters degree in biology at Texas Tech University in Lubbock, Texas, in the laboratory of Llewellyn Densmore, where he focused on molecular ecology of water snakes and crocodiles.

In 2007, John accepted a position in the laboratory of Christopher Austin at LSU. The day he moved, he met his future wife, Megan Apperson, as he was moving into his new apartment (she was moving out). One year later he began his doctoral study with Bryan Carstens at LSU. John currently resides with his wife in Durham, NC.